Statistical methods and software for clinical trials with binary and survival endpoints

Efficiency, sample size and two-sample comparison

Marta Bofill Roig

PhD Thesis directed by:
Guadalupe Gómez Melis

Thesis submitted to obtain the title of Doctor
by the Universitat Politècnica de Catalunya.

Department of Statistics and Operations Research
Barcelona, November 14, 2020

This research was supported by: Ministerio de Economía y Competitividad (Spain) under Grant MDM-2014-0445, through the María de Maeztu Programme for Units of Excellence in R&D, and under Grant MTM2015-64465-C2-1-R (MINECO/FEDER, UE); Ministerio de Ciencia e Innovación (Spain) under Grant PID2019-104830RB-I00; Departament d'Empresa i Coneixement de la Generalitat de Catalunya (Spain) under Grant 2017 SGR 622 (GRBIO).

A Rosa M. Bofill Serra

Acknowledgements

A la meva directora i mentora, Lupe, que m'ha guiat i ajudat durant tota aquesta etapa; que m'ha ensenyat a estimar la recerca i transmès la passió i la dedicació cap a aquesta; que m'ha animat quan estava desanimada, i recolzat i empès a perseguir totes les noves idees i ambicions. Per tot el que hem compartit durant aquests anys i per tot el que ens queda per recórrer plegades, moltes gràcies.

A totes i tots els membres del GRBIO, que des de l'inici m'han oferit la seva ajuda i dedicat el seu temps en aconsellar-me. En especial, a l'Alex, al Klaus, al Jordi i a la Núria que m'han ajudat sempre que m'ha calgut i amb els qui he gaudit molt treballant aquests anys.

Als meus companys de la UPC, Cecilia, David i Yovaninna, amb qui vaig tenir la sort de començar el doctorat; que m'han escoltat assajar (i reassajar) totes les presentacions i que han vist, dia a dia, créixer aquest projecte, de la mateixa manera que jo he vist créixer els seus. Al Pol i al Guille perquè plegats ens hem entusiasmat parlant d'estadística, animat davant dels reptes, i apuntat a totes les iniciatives i activitats que se'ns han posat al davant. A l'Àlvar, l'Evi i la Lore amb qui després d'un breu pas pel departament m'uneix una gran amistat.

Durant aquests anys, he participat en moltes de les activitats, seminaris i congressos organitzats per la SEB, SoCE i BIOSTATNET. Gràcies a això, he tingut la sort de conèixer a un bon grapat de persones a qui tinc molta estima i admiració. Vull agrair especialment el carinyo i els ànims de: Antonio Martín, Carles Serrat, Coté Rodríguez-Álvarez, Inma Arostegui, Malu Calle, Montse Rue, Núria Porta, Oleguer Plana, Pere Puig i Rosa Lamarca. I el meu record per l'Anna Espinal.

Gràcies també al grup de joves i amics de la JJSEB, en particular, a: Amaia Iparragirre, Irantzu Barrio, Joaquín Martínez, Josu Najera-Zuloaga i Natàlia Vilor.

I would also like to thank the Department of Biostatistics at the University of Texas MD Anderson Cancer Center for the opportunity to visit them. I am

specially grateful to Prof. Yu Shen for her kind help. Thanks also to Prof. Martin Posch for my research stay at the Section of Medical Statistics at the Medical University of Vienna. It has been a great pleasure to work with you, and I hope to keep working together in the near future.

Als meus amics, Elena, Juanfran, Maria, Javi i Mònica, per tots els ànims, per entendre'm i per treure'm de la bombolla de la tesi quan més ho necessitava.

A la meva família per tot el seu suport. Especialment, als meus avis i a l'Andrea, la Sandra i la Paula. A la meva germana, Ester, que m'ha vist emocionar-me, patir, plorar i riure amb aquesta tesi; i que m'ha escoltat dia sí dia també parlar de la meva feina. Per aquests anys, vivint i compartint tot plegades, que he disfrutat i que tornaria a repetir. Als meus pares, que sempre m'han animat a perseguir els meus somnis, que m'han inculcat el valor de l'esforç, recolzat en les decisions i ajudat en els moments difícils.

A totes i a tots.
Gràcies,

Marta.

Abstract

Defining the scientific question is the starting point for any clinical study. However, even though the main objective is generally clear, how this is addressed is not usually straightforward. Clinical studies very often encompass several questions, defined as primary and secondary hypotheses, and measured through different endpoints.

In clinical trials with multiple endpoints, composite endpoints, defined as the union of several endpoints, are widely used as primary endpoints. The use of composite endpoints is mainly motivated because they are expected to increase the number of observed events and to capture more information than by only considering one endpoint. Besides, it is generally thought that the power of the study will increase if using composite endpoints and that the treatment effect on the composite endpoint will be similar to the average effect of its components. However, these assertions are not necessarily true and the design of a trial with a composite endpoint might be difficult.

Different types of endpoints might be chosen for different research stages. This is the case for cancer trials, where short-term binary endpoints based on the tumor response are common in early-phase trials, whereas overall survival is the gold standard in late-phase trials. In the recent years, there has been a growing interest in designing seamless trials with both early response outcome and later event times. Considering these two endpoints together could provide a wider characterization of the treatment effect and also may reduce the duration of clinical trials and their costs.

In this thesis, we provide novel methodologies to design clinical trials with composite binary endpoints and to compare two treatment groups based on binary and time-to-event endpoints. In addition, we present the implementation of the methodologies by means of different statistical tools. Specifically, in Chapter 2, we propose a general strategy for sizing a trial with a composite binary endpoint as

primary endpoint based on previous information on its components. In Chapter 3, we present the ARE (Asymptotic Relative Efficiency) method to choose between a composite binary endpoint or one of its components as the primary endpoint of a trial. In Chapter 4, we propose a class of two-sample nonparametric statistics for testing the equality of proportions and the equality of survival functions. In Chapter 5, we describe the software developed to implement the methods proposed in this thesis. In particular, we present CompARE, a web-based tool for designing clinical trials with composite endpoints and its corresponding R package, and the R package `SurvBin` in which we have implemented the class of statistics presented in Chapter 4. We conclude this dissertation with general conclusions and some directions for future research in Chapter 6.

Resumen

La evaluación de la eficacia de los tratamientos es uno de los mayores retos en el diseño de ensayos clínicos. La variable principal cuantifica la respuesta clínica y define, en gran medida, el ensayo. Los ensayos clínicos generalmente abarcan varias cuestiones de interés. En estos casos, se establecen hipótesis primarias y secundarias, que son evaluadas a través de diferentes variables.

Los ensayos clínicos con múltiples variables de interés utilizan frecuentemente las llamadas variables compuestas. Una variable compuesta se define como la unión de diversas variables de interés. La utilización de variables compuestas en lugar de variables simples estriba en que con éstas aumenta el número de eventos observados y se obtiene una información más completa sobre la respuesta al tratamiento. También se plantea a menudo, por un lado, que la potencia estadística del estudio es mayor si se usan variables compuestas y, por otro, que el efecto del tratamiento de la variable compuesta será similar al efecto medio de las variables que la componen. Sin embargo, estas afirmaciones no son necesariamente ciertas y el diseño de un estudio con una variable compuesta suele ser complejo.

El tipo de variable escogida como variable principal puede diferir en las diferentes etapas de investigación. Por ejemplo, en el caso de estudios oncológicos, las variables binarias evaluadas a corto plazo son usadas en fases tempranas del desarrollo del tratamiento; mientras que en fases más avanzadas, las variables más usadas son tiempos de vida. En los últimos años, ha habido un interés creciente en el diseño de ensayos fase II/III con variables binarias y tiempos de vida. Este tipo de ensayos podría proporcionar una caracterización más amplia del efecto del tratamiento y también podría reducir la duración de los ensayos clínicos y sus costes.

En esta tesis, proponemos nuevas metodologías, junto con el software estadístico correspondiente, para el diseño de ensayos clínicos con variables compuestas y para

la comparación de dos grupos de tratamiento en base a variables binarias y tiempos de vida. Específicamente, en el capítulo 2, proponemos una estrategia para calcular el tamaño muestral de un ensayo con una variable compuesta como variable principal del estudio basado en la información previa sobre sus componentes. En el capítulo 3, presentamos el método ARE (*Asymptotic Relative Efficiency*) para elegir entre una variable compuesta o una de sus componentes como variable principal de un ensayo. En el capítulo 4, proponemos una clase de estadísticos no paramétricos para contrastar la igualdad de proporciones y la igualdad de las funciones de supervivencia. En el capítulo 5, describimos el software desarrollado para implementar los métodos propuestos en esta tesis. En particular, presentamos CompARE, una herramienta web para diseñar ensayos clínicos con variables compuestas y su correspondiente paquete R, y el paquete R `SurvBin` en el que hemos implementado la clase de estadísticos presentadas en el capítulo 4. La tesis concluye con un resumen de las principales aportaciones, algunas conclusiones de carácter general así como con una discusión sobre diversos problemas abiertos y futuras líneas de investigación.

Resum

L'avaluació de l'eficàcia dels tractaments és un dels grans reptes en el disseny d'assajos clínics. La variable principal quantifica la resposta clínica i defineix, en gran manera, l'assaig. Els assaigs clínics generalment inclouen diverses qüestions d'interès. En aquests casos, s'estableixen hipòtesis primàries i secundàries, que són avaluades mitjançant diferents variables.

Els assajos clínics amb múltiples variables d'interès utilitzen freqüentment les anomenades variables compostes. Una variable composta es defineix com la unió de diverses variables d'interès. La utilització de variables compostes en lloc de variables simples rau en el fet que amb aquestes augmenta el nombre d'esdeveniments observats i s'obté una informació més completa sobre la resposta al tractament. També es planteja sovint, d'una banda, que la potència estadística de l'estudi és més gran si es fan servir variables compostes i, de l'altra, que l'efecte del tractament de la variable composta serà semblant a l'efecte mitjà de les variables que la composen. No obstant això, aquestes afirmacions no són necessàriament certes i el disseny d'un estudi amb una variable composta sol ser complex.

El tipus de variable escollida com a variable principal pot diferir en les diferents etapes d'investigació. Per exemple, en el cas d'estudis oncològics, les variables binàries avaluades a curt termini són utilitzades en fases inicials; mentre que en fases més avançades, les variables més utilitzades són temps de vida. En els últims anys, hi ha hagut un interès creixent en el disseny d'assaigs fase II/III amb variables binàries i temps de vida. Aquest tipus d'assajos podria proporcionar una caracterització més àmplia de l'efecte del tractament i també podria reduir la durada dels assaigs clínics i els seus costos.

En aquesta tesi, proposem noves metodologies, juntament amb el software estadístic corresponent, per al disseny d'assajos clínics amb variables compostes i per a la comparació de dos grups de tractament a partir de variables binàries i temps de vida. Específicament, en el capítol 2, proposem una estratègia per

calcular la mida mostral d'un assaig amb una variable composta com a variable principal d'estudi basat en la informació prèvia sobre els seus components. En el capítol 3, presentem el mètode ARE (*Asymptotic Relative Efficiency*) per triar entre una variable composta o una de les seves components com a variable principal d'un assaig. En el capítol 4, proposem una classe d'estadístics no paramètrics per contrastar la igualtat de proporcions i la igualtat de les funcions de supervivència. En el capítol 5, descrivim el software desenvolupat per implementar els mètodes proposats en aquesta tesi. En particular, presentem CompARE, una eina web per dissenyar assajos clínics amb variables compostes i el seu corresponent paquet d'R, i el paquet d'R `SurvBin` on hem implementat la classe d'estadístics presentada en el capítol 4. La tesi conclou amb un resum de les principals aportacions, algunes conclusions de caràcter general així com amb una discussió sobre diversos problemes oberts i futures línies d'investigació.

Contents

Chapter 1
Introduction

Statistics come along with scientists in all stages of investigation of clinical research, from design to analysis and prediction. Clinical research is most times focused on the prevention or treatment of diseases, drug discovery being the cornerstone of medical progress. Medical advances move forward through experimentation and by using several study designs. There are two main types of clinical studies depending on the study design and thus on the nature of data: clinical trials and observational studies (Cook et al., 2008).

When a new drug is being investigated, the usual method for studying its efficacy is by comparing how it acts with respect to a placebo or standard of care (Pocock, 1983). The groups of patients that receive either of the treatments must be as similar as possible and should only differ in the treatment that each group receives. Otherwise, observed differences between groups may not be attributable to the treatments, but can arise from other characteristics of the groups.

In clinical trials, the assignment of subjects to a group is through a randomization process. Randomization is, in its most simple form, a process by which all subjects are equally likely to be assigned to either one of the treatment groups (Friedman et al., 2010). When data come from clinical trials, we are able to answer questions about the efficacy of a treatment with respect to another, or in other words, to state causal relationships between the treatment and the response. In observational studies, however, there is no randomly assignment of subjects to groups. Such studies are mostly focused on the associational relationships between the treatment and the outcome and can not generally draw causal inferences between them due to potential biases.

The emergence of clinical trials in 1948 laid the foundation for modern drug development process (Medical Research Council, 1948). Subsequent advances in ethics, statistical methodologies, and protocols and regulatory issues have led to the development and improvement of clinical trials since their origin. But still,

clinical trials are not exempted of challenges and they are continuously growing and adapting to new problems.

Clinical trials are considered the gold standard procedure to evaluate efficacy and safety of a new drug in human beings. However, the process until there is a new drug to be tested is lengthy, and clinical trials are not just limited to comparative studies when a candidate drug is available. Clinical trials put together several steps, or phases, of drug discovery and are commonly classified into four phases, each of them following a different purpose and helping scientists to answer different questions (Friedman et al., 2010; Senn, 2008).

In phase I trials, the new drug is usually tested in healthy volunteers aiming to estimate its tolerability and to characterize how the drug affects the organism and vice versa. The main statistical challenges in phase I trials are to relate the dose to the toxicity of the new drug and to establish the maximum tolerated dose in a small group of patients (Storer, 1989; Buoen et al., 2005). Once the range of appropriate doses is determined, in phase II trials the drug is tested in a larger group of diseased patients with the objective of evaluating its efficacy. There are some phase II studies that, because of uncertainty in dose-response, consider several doses; whereas others use a fixed dose chosen based on previous phase I trial results. The goals in this phase are identifying the optimal dose level and estimating the effect of that dose on diseased patients. It is common to use multiple testing procedures in order to acquire the optimal dose level. Moreover, adaptive designs containing stopping rules are often considered for cases where the drug is either very toxic or very effective, allowing the trial to be stopped early. Phase II trials do not usually employ a comparative design. During this phase, the treatment is usually evaluated on endpoints that can be measured in a short period of time and that are related to the clinical outcome. The primary objective of phase II trials is to determine whether the new drug should be used in a large-scale comparative trial and, if so, to estimate the response so that it can aid investigators in designing further studies. How precise the estimated response is will be crucial in the design of subsequent controlled trials.

Phase III trials have much larger sample sizes, recruiting hundreds or even thousands of patients, and aim to compare the new drug against the standard of care. These trials are carefully designed, in concordance with the regulatory guidelines, to ensure the reliability of the trial and the drug's efficacy estimation, to monitor side-effects, and to compare it to already approved treatments. Many statistical challenges arise in this stage such as the selection and analysis of appropriate

endpoints, or the calculation of the required sample size for detecting an specific effect size.

Phase IV trials are conducted after a drug has been already approved by the regulatory agencies. The purpose of these trials is to learn more about the side effects and safety of the drug, or to assess long term risks and benefits in a broad, general population not subjected to strict clinical scrutiny.

This classification is very general and might be very flexible. Additionally, there may be differences depending on the disease under study. Whereas early phase studies may be controlled or uncontrolled, later phases are usually controlled trials with longer follow-up periods and with larger sample sizes. Not all the trials may fit into a single phase, there are some blending different phase studies, for instance, phase I/II trials. In the recent years, there has been an increasing number of hybrids of phase II and phase III trials in order to foster a faster and more efficient drug development process (Lai et al., 2012; Thall et al., 2012; Kieser et al., 2018). Also the large number of promising drugs that fail in phase III has lead to enlarge phase II trials to guarantee that the findings are clinically relevant and to improve the estimates of the effect sizes (De Martini, 2019) to be used in later phases.

Clinical trial designs often encounter the need of using more than one event to measure the efficacy of treatment effect, specially in phase III trials. The use of multiple responses contributes to a wider picture of the intervention effects providing more information. Trials involving multiple endpoints commonly arise in practice. For instance, in cardiovascular studies, multiple endpoints are often considered, such as myocardial infarction, acute coronary syndrome, stroke and death.

Multiple endpoints put forth challenges in the design and analysis of clinical trials. How we deal with multiple endpoints will rely on the question that it is addressed, since the efficacy claim in this situation might be defined in different ways. The efficacy endpoints are called co-primary endpoints when it is necessary to demonstrate a treatment effect on each of the endpoints to conclude that the drug is effective; while they are called multiple primary endpoints when it is sufficient to demonstrate a treatment effect on at least one of the endpoints to conclude on the effectiveness of the drug (EMA Guideline, 2017). When designing the trial to evaluate the effect on co-primary endpoints, no adjustment is needed to control the type I error, but the statistical power decreases as the number of endpoints to be evaluated increases (FDA Guidance, 2017). On the other hand,

when designing the trial to evaluate an effect on multiple primary endpoints, an adjustment is needed to control the type I error rate.

One popular alternative to multiple endpoints is to reduce the multi-dimensional problem into a one-dimensional problem by collapsing the information of several responses into a single endpoint. In the context of survival and binary data, composite endpoints, defined as the union of several events, are frequently applied. In this case, treatment effect is evaluated on the time until the occurrence of the first of several events; or on the binary endpoint that takes value 1 whenever one of the outcomes has occurred.

The growth in the number of trials with multiple endpoints has contributed to the need of further research in the design and analysis of such trials. One major concern is how to determine the required sample size. As it is well-known, if the sample size is less than necessary the trial might not be able to conclude meaningful results. On the other hand, the use of more sample size than necessary might detect treatment differences too small to be clinically relevant. Both cases would imply a waste of resources and an unnecessary patients' risk exposure (Cook et al., 2008).

The sample size calculation when multiple endpoints are involved mainly relies upon the clinical question, which in turn defines the alternative hypothesis to be tested. Many authors have proposed different approaches to size trials with multiple primary and co-primary endpoints. Sozu et al. (2010) and Sozu et al. (2015) discussed sample size formulae for multiple co-primary and multiple primary binary endpoints, respectively; Senn and Bretz (2007) discussed sample size for trials with multiple co-primary and multiple primary continuous endpoints; and Sugimoto et al. (2017) presented sample size calculations for multiple co-primary and multiple primary time-to-event endpoints. Surprisingly, less attention has been given to the design and sample size with composite endpoints.

Sometimes the major difficulty in the sample size calculation is that the required information depends on parameters which are often unknown or highly variable. In trials with multiple endpoints, the association among the considered endpoints should be taken into account to obtain the appropriate sample size (FDA Guidance, 2017). However, this association is usually unknown and difficult to obtain. Although several authors have assessed how the degree of association between endpoints affects on sample size when using multiple co-primary binary endpoints (Ando et al., 2015; Sozu et al., 2010), methodologies for sample size calculations which address how to deal with the lack of knowledge of the correlation are limited.

Outline of the thesis

This thesis deals with the design and analysis of late phase trials with multiple endpoints and it consists of two different parts. The first part concerns the design of trials with composite binary endpoints; whereas in the second a binary endpoint and a time-to-event endpoint are considered for the comparison of two treatment groups.

Composite binary endpoints are commonly used as primary endpoints in clinical trials. When designing a trial, it is crucial to determine the appropriate sample size for testing the statistical differences between treatment groups for the primary endpoint. As shown in Chapter 2, when using a composite binary endpoint to size a trial, one needs to specify the event rates and the effect sizes of the composite components as well as the correlation between them. In practice, the marginal parameters of the components can be obtained from previous studies or pilot trials; however, the correlation is often not previously reported and thus usually unknown. In Chapter 2, we first show that the sample size for composite binary endpoints is strongly dependent on the correlation and, second, that slight deviations in the prior information on the marginal parameters may result in underpowered trials for achieving the study objectives at a pre-specified significance level. We propose a general strategy for calculating the required sample size when the correlation is not specified and accounting for uncertainty in the marginal parameter values. Chapter 2 mainly reproduces the publication:

A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components.
Bofill Roig, M., and Gómez Melis, G.
Statistics in Medicine. Volume 38, Issue 11, 20 May 2019, Pages 1935–1956.
DOI: 10.1002/sim.8092.

Chapter 3 focuses on the choice of the primary endpoint in trials with composite endpoints. This choice is an important issue when designing a clinical trial. It is common to use composite endpoints as a primary endpoint because it increases the number of observed events, captures more information and is expected to increase the power. However, combining events that have no similar clinical importance and have different treatment effects makes the interpretation of the results cumbersome and might reduce the power of the corresponding tests. Gómez and Lagakos

(2013) proposed the Asymptotic Relative Efficiency (ARE) method to choose between a composite or one of its components as primary endpoint for comparing the efficacy of a treatment based on the times to each of these endpoints. In Chapter 3, we expand the ARE method to binary endpoints. We show that the ARE method depends on six parameters including the degree of association between components, event proportion, and effect of therapy given by the corresponding single endpoints. The main content of Chapter 3 has been published in:

Selection of composite binary endpoints in clinical trials.
Bofill Roig, M., and Gómez Melis, G.
Biometrical Journal. Volume 60, Issue 2, March 2018, Pages 246-261.
DOI: 10.1002/bimj.201600229.

Lifetime analysis has often been the sharp focus of clinical trial research. In trials with multiple endpoints, the time until the event is not always the outcome of interest for all endpoints while the occurrence of an event over a fixed time period is important in itself. In cancer immunotherapies trials, short-term binary endpoints based on the tumor size, such as objective response, are common in early-phase trials, whereas overall survival remains the gold standard in late-phase trials (Ananthakrishnan and Menon, 2013). In Chapter 4, we propose a class of two-sample statistics for testing the equality of proportions and the equality of survival functions. We build our proposal on a weighted combination of a score test for the difference in proportions and a Weighted Kaplan-Meier statistic-based test for the difference of survival functions. The proposed statistics are fully nonparametric and do not rely on the proportional hazards assumption for the survival outcome. We present the asymptotic distribution of these statistics, propose a variance estimator and show their asymptotic properties under fixed and local alternatives. The proposed class of statistics could be used in seamless phase II/III cancer trials, where our approach would provide an insight of the tumor activity through the binary endpoint while the study continues with the time-to-event response. The work presented in Chapter 4 reproduces the following paper currently under review:

A class of two-sample nonparametric statistics for binary and time-to-event outcomes.
Bofill Roig, M., and Gómez Melis, G.
arXiv:2002.01369 [stat.ME]

Chapter 5 contains a description of the software developed to implement the methodologies presented in this thesis. Specifically, we present CompARE, a web-

based tool for designing clinical trials with composite endpoints and its corresponding R package, and the R package `SurvBin` in which we have implemented the class of statistics presented in Chapter 4. We end the thesis with some final conclusions and possible lines for future research in Chapter 6.

Chapter 2

Sample size for composite binary endpoints

The main content of this chapter has been published in:

A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components.

Bofill Roig, M., and Gómez Melis, G.

Statistics in Medicine. Volume 38, Issue 11, 20 May 2019, Pages 1935–1956.

DOI: 10.1002/sim.8092.

Sections 2.1 to 2.7 reproduce almost identically the corresponding work in the paper. However some modifications have been done for coherence and cohesion of the thesis. The illustration in Section 2.5 has been analysed using the software CompARE, though we postpone the description of CompARE to Chapter 5. Section 2.8 summarizes the contents of the online supplementary material of the paper, and Section 2.9 describes further work that has not been published. The discussion corresponds mostly to the one in the paper, but we have slightly changed some paragraphs to update and encompass the further work we have made. Proofs and technical details are in the Appendix A.

2.1 Introduction

Many trials are designed to evaluate more than one endpoint with the aim of providing a wider picture of the intervention effects (FDA Guidance, 2017; Rosenblatt, 2017). When the rate of occurrence of an event is expected to be low, it is common to consider the composite event defined as the occurrence of any of a set of pre-specified events. This composite event is usually chosen as the primary efficacy endpoint for comparing two treatment groups, either by comparing proportions between groups at the end of the study or by using time-to-event analysis. In this work, we focus on composite binary endpoints.

Power analysis and its subsequent sample size calculation have been widely discussed in the literature on comparing two proportions in the univariate case (Lachin, 1981; Donner, 1984; Fleiss, 1981). These standard sample size formulae are based on the effect size and the frequency of occurrence of primary endpoint, and they could be applied in a straightforward way to a composite endpoint if its effect size and frequency are known prior to the initiation of the study. However, the effect size and frequency of observing the composite endpoint depend on the corresponding effect and frequency of the composite components, which are often quite dissimilar and thus make the composite parameters very difficult to anticipate.

The TACTICS-TIMI 18 trial (Cannon et al., 2001) illustrates some problems that might arise when determining the sample size for a primary composite binary endpoint. TACTICS-TIMI 18 was an international, multicenter, randomized trial that evaluated the efficacy of invasive and conservative treatment strategies in patients with unstable angina or non-Q-wave acute myocardial infarction treated with tirofiban, heparin, and aspirin. The primary hypothesis of the TACTICS-TIMI 18 trial was that an early invasive strategy would reduce the combined incidence of death, acute myocardial infarction, and rehospitalization for acute coronary syndromes at six months when compared with an early conservative strategy. The primary endpoint was the composite endpoint formed by death, non-fatal myocardial infarction, and rehospitalization for acute coronary syndrome at 6 months.

Similar research questions such as those in TACTICS-TIMI 18 were previously investigated in the TIMI IIIB and VANQWISH trials (Cannon et al., 1998). The TIMI IIIB trial (Anderson et al., 1995) considered the primary composite endpoint of death, post-randomization myocardial infarction, and a positive exercise test at 6 weeks; whereas the primary endpoint in the VANQWISH trial (Boden et al., 1998) was the combination of death and non-fatal myocardial infarction

at 12 months of follow-up. The initial planning of TACTICS-TIMI 18 was based on those trials expecting 22% events of the primary composite endpoint in the conservative-strategy group, to detect a relative difference of 25% between the two groups for a 80% power. Those anticipated values resulted in the need to recruit at least 1720 patients. However, TACTICS-TIMI 18 yielded a 19% frequency of observing the combination of death, acute myocardial infarction and rehospitalization at six months, which was remarkably lower than expected and delivered a relative difference of 20% between groups, a figure that is seriously lower than the anticipated 25%. Note that if the anticipated frequency of observing the composite endpoint had been closer to the observed results, at least 2000 patients rather than 1720 would have been required and the sample size needed would have been larger than the one initially planned.

In this chapter, we present sample size formulations for detecting a hypothesized difference between treatments in a primary composite binary endpoint based on the event rates and effect sizes of the composite components. The motivation for this is mainly because prior information on the marginal effects and event rates is commonly available from previous or pivotal studies, as illustrated in the TACTICS-TIMI 18 trial. Moreover, the major findings in a trial with a primary composite endpoint should be well supported by its components (FDA Guidance, 2017; EMA Guideline, 2017), since the trial could be considered negative if the components are not in line with the result (Pocock et al., 2015; ICH9, 1999). Nevertheless, as shown in this work, the sample size calculation for composite endpoints relies not only on the anticipation of the effect size and the event rates of the composite components, but also on the correlation between them. However, even though the marginal parameters could be obtained previously, the correlation is usually not reported in practice and, thus, is frequently unknown and difficult to anticipate.

Several authors have addressed the correlation's influence on sample size determination when more than one endpoint is used as the primary endpoint. Sozu et al. (2010) discuss several methods for calculating power and sample size for multiple co-primary binary endpoints, and they study the impact on the sample size, specifically regarding the association among endpoints. Senn and Bretz (2007) examine sample size for trials under different power definitions for multiple testing problems. Rauch and Kieser (2012) and Sander et al. (2016) define a multiple test procedure focused on a composite binary endpoint and a pre-specified main component, and propose an internal pilot study for estimating the unknown parameters and revising the sample size. However, to the best of our knowledge,

methodologies are limited in regard to handling the sample size calculation for composite binary endpoints when the correlation is unknown.

In this work, we focus on providing a general procedure for sizing trials with composite binary endpoints, doing so on the basis of anticipated information of the composite components even if the correlation is unknown. We show that the sample size for composite binary endpoints is strongly dependent on the correlation, and that slight deviations in the prior information on the marginal parameters may result in trials being too underpowered for achieving the study objectives at the pre-specified significance level. We propose a sample size strategy to calculate the minimum sample size that guarantees the planned power while accounting for, on the one hand, the uncertainty of the correlation value and, on the other, plausible deviations in the marginal parameter values. Furthermore, we have implemented the methodologies presented in this chapter in CompARE. CompARE is a freely available web-based tool for characterizing binary composite endpoints and computing the needed sample size under several settings. CompARE provides aids to help understand the role played by each one of the components of the composite endpoint, as well as their consequences on the required sample size. In this chapter, we use CompARE to illustrate the proposed methods. The presentation of CompARE is postponed to Chapter 5.

This chapter is structured as follows. In Section 2.2, we introduce the settings of the problem. In Section 2.3, we review sample size planning when evaluating risk difference. In Section 2.4, we present sample size formulae for composite binary endpoints based on the parameters of the components plus the correlation. We further describe the performance of these formulae according to the parameters and propose a strategy for sizing trials when the correlation is unknown. In Section 2.5, we exemplify the proposal by the TACTICS-TIMI 18 trial using CompARE, and in Section 2.6 we extend the proposal to those trials for which the treatment effect is measured by the relative risk or odds ratio. In Section 2.7, we investigate the performance of the power and significance level under misspecification of the correlation and evaluate the proposed sample size strategy with a simulation study. In Section 2.8, we outline the contents of the supplementary material of Bofill and Gómez (2019). In Section 2.9 we describe different association measures for binary endpoints and state some properties of the composite effect. We conclude the chapter with the Discussion.

2.2 Notation and assumptions

We consider a randomized clinical trial comparing two treatment groups: the control group ($i = 0$) and treatment group ($i = 1$), each one composed of $n^{(i)}$ patients who are followed for a pre-specified time τ. For simplicity, we consider only two events of potential interest, ε_1 and ε_2. Let X_{ijk} denote the response of the k-th binary endpoint for the j-th patient in the i-th group of treatment ($i = 0, 1$, $j = 1, ..., n^{(i)}$, $k = 1, 2$). The response X_{ijk} is defined by 1 if the event, ε_k, has occurred during the follow-up and 0 otherwise.

We define the binary composite endpoint as the event that occurs whenever one of the endpoints is observed, that is, $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$. At this point we assume that the composite endpoint is well-defined, that is, both composite components are important enough to be considered; and we include those adverse clinical outcomes that are relevant to the clinical setting. We denote by X_{ij*} the composite response defined as a Bernoulli random variable with probability of observing the event $p_*^{(i)} = \mathrm{P}(X_{ij*} = 1) = 1 - q_*^{(i)}$, where:

$$X_{ij*} = \begin{cases} 1, \text{ if } X_{ij1} + X_{ij2} \geq 1 \\ 0, \text{ if else } X_{ij1} + X_{ij2} = 0 \end{cases} \tag{2.1}$$

To evaluate whether there is a risk reduction in the treatment group compared with the control group, we set a hypothesis test where the null hypothesis states that there is no difference between the control and the treatment groups; whereas the alternative hypothesis assumes a risk reduction in the treatment group. The usual measures to evaluate the treatment effect when comparing two groups are the difference in proportions (also called risk difference), denoted by δ_*; the relative risk (or risk ratio), R_*; and the odds ratio, OR_*. The relationship between these measures and the probabilities of observing the binary composite endpoint in each group are given in Table 2.1, together with the null and alternative hypothesis that should be set in each case. The following sections will be developed in terms of the risk difference $\delta_* = p_*^{(1)} - p_*^{(0)}$ of the composite binary endpoint. Section 2.6 extends the results to the relative risk and odds ratio.

2.2.1 An insight into the parameters of the composite endpoint

Let $p_k^{(i)}$ and $q_k^{(i)}$ represent the probabilities that ε_k occurs or not, respectively, for a patient belonging to the i-th group. Let $\rho^{(i)}$ denote Pearson's correlation coeffi-

Table 2.1 Parameter to anticipate the effect, and set of hypotheses.

	Parameter effect	**Null hypothesis**	**Alternative hypothesis**
Risk difference	$\delta_* = p_*^{(1)} - p_*^{(0)}$	$\delta_* = 0$	$\delta_* < 0$
Relative risk	$\mathrm{R}_* = p_*^{(1)}/p_*^{(0)}$	$\log(\mathrm{R}_*) = 0$	$\log(\mathrm{R}_*) < 0$
Odds ratio	$\mathrm{OR}_* = \frac{p_*^{(1)}/q_*^{(1)}}{p_*^{(0)}/q_*^{(0)}}$	$\log(\mathrm{OR}_*) = 0$	$\log(\mathrm{OR}_*) < 0$

cient between the components in the i-th group. The probability of observing the composite event ε_* is in terms of the probabilities of ε_1 and ε_2 and the correlation, as follows:

$$p_*^{(i)} = 1 - q_1^{(i)} q_2^{(i)} - \rho^{(i)} \sqrt{p_1^{(i)} p_2^{(i)} q_1^{(i)} q_2^{(i)}}\ , \qquad i = 0, 1 \tag{2.2}$$

Note here that the probability of observing the composite endpoint becomes smaller as the correlation between the components of the composite increases.

The effect size in the composite endpoint in terms of the risk difference, δ_*, is given by:

$$\begin{aligned} \delta_* = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1 \delta_2 + \rho^{(0)} \sqrt{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}} \\ -\rho^{(1)} \sqrt{(p_1^{(0)} + \delta_1)(p_2^{(0)} + \delta_2)(q_1^{(0)} - \delta_1)(q_2^{(0)} - \delta_2)} \end{aligned} \tag{2.3}$$

where δ_k $(k = 1, 2)$ corresponds to the risk difference for each of its components.

From now on the correlation is assumed equal for both groups and denoted by ρ, that is, $\rho = \rho^{(0)} = \rho^{(1)}$. A short discussion on the consequences on the assumption $\rho^{(0)} = \rho^{(1)}$ is available in the supplementary material (see Section 2.8). Let θ denote the vector of event rates of the composite components in the control group, that is, $\theta = (p_1^{(0)}, p_2^{(0)})$, and let λ represent the vector of marginal effect sizes, that is, $\lambda = (\delta_1, \delta_2)$. We will denote the risk difference as a function of the marginal parameters (θ, λ) and the correlation ρ by $\delta_*(\theta, \lambda, \rho)$; and the probability of observing ε_* under the control group by $p_*^{(0)}(\theta, \rho)$. We remark here that when λ and θ are fixed such that $p_k^{(0)} < 0.5$ and $\delta_k < 0$ $(k = 1, 2)$, the risk difference $\delta_*(\theta, \lambda, \rho)$ increases with respect to the correlation ρ (see Appendix A.1).

2.3 Sample size when the parameters of the composite endpoint can be anticipated

In this section we summarize the statistics and sample size formulae to test for a risk difference when the probability of occurrence in the control group of the composite binary endpoint can be anticipated and for a given expected risk difference. Since the composite endpoint is an univariate outcome, a single statistical test is performed and, consequently, no multiplicity problem occurs and no statistical adjustment is needed. Therefore, as we will see, the formulas follow the univariate case and are straightforward but to make this work comprehensive and the following sections meaningful, we displayed them in terms of the composite endpoint parameters.

Herein, we assume a clinical trial where, first, patients are randomized to one of two treatment arms following a balanced design and, second, where the primary endpoint is a binary composite endpoint. The aim is to detect a hypothesized risk reduction in the primary composite endpoint at the significance level of α and with desired power equal to $1-\beta$. Let n be the total sample size required, with $n^{(i)} = n/2$ patients per group ($i = 0, 1$); and let us denote by z_α and z_β the values of standardized normal deviates corresponding to α and β.

The null hypothesis is stated as $\mathrm{H}_0^* : p_*^{(1)} - p_*^{(0)} = 0$ and is compared against the alternative hypothesis $\mathrm{H}_1^* : p_*^{(1)} - p_*^{(0)} < 0$. To test H_0^* against H_1^* we use the statistic:

$$\mathrm{T}_{*,n} = \frac{\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}}{\sqrt{\widehat{Var}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)})}} \tag{2.4}$$

where $\widehat{p}_*^{(i)} = \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} X_{ij*}$. Under H_0^*, $\mathrm{T}_{*,n}$ follows, asymptotically, the standard normal distribution. We will reject the null hypothesis at the α level of significance if $\mathrm{T}_{*,n} < -z_\alpha$. The variance $Var(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)})$ in equation (2.4) can be estimated under H_0^* using the pooled variance estimate (Donner, 1984):

$$\widehat{Var}_{H_0}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}) = \frac{1}{2n^{(0)}} \cdot \left(\widehat{p}_*^{(0)} + \widehat{p}_*^{(1)}\right) \cdot \left(\widehat{q}_*^{(0)} + \widehat{q}_*^{(1)}\right)$$

or under H_1^* using the unpooled variance estimate:

$$\widehat{Var}_{H_1}(\widehat{p}_*^{(1)} - \widehat{p}_*^{(0)}) = \frac{1}{n^{(0)}} \left(\widehat{p}_*^{(0)}\widehat{q}_*^{(0)} + \widehat{p}_*^{(1)}\widehat{q}_*^{(1)}\right)$$

For a given probability under control group $p_*^{(0)}$, the required sample size using the pooled estimate to have power $1-\beta$ in order to detect an effect size of δ_* at a significance level α is given by (Lachin, 1981; Fleiss, 1981):

$$n = 2 \cdot \left(z_\alpha \cdot \sqrt{(2p_*^{(0)} + \delta_*)(2q_*^{(0)} - \delta_*)} + z_\beta \cdot \sqrt{p_*^{(0)} q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*)} \right)^2 \Big/ \delta_*^2 \tag{2.5}$$

Note that in (2.5) we have replaced $p_*^{(1)}$ with $p_*^{(0)} + \delta_*$.

Similarly, the corresponding sample size using the unpooled variance estimate is given by:

$$n = 2 \cdot \left(\frac{z_\alpha + z_\beta}{\delta_*} \right)^2 \cdot \left(p_*^{(0)} q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*) \right) \tag{2.6}$$

Note that, under the null hypothesis $\mathrm{H}_0^* : p_*^{(1)} - p_*^{(0)} = 0$, expressions (2.5) and (2.6) coincide.

2.4 Sample size based on anticipated values of the composite components

Sample size formulae underlined in Section 2.3 are based on the parameters of the composite endpoint, that is, the event rate under the control group, $p_*^{(0)}$, and the treatment effect, δ_*. In this section, we derive the sample size based on the anticipated information on the marginal parameter values and the correlation, even if the correlation value is not fully specified and/or the event rates values are not accurately anticipated.

2.4.1 Sample size based on composite components

Given the event rates in the control group $\theta = (p_1^{(0)}, p_2^{(0)})$, the expected effect size for each component $\lambda = (\delta_1, \delta_2)$, and the correlation between the occurrence of both components ρ, we will denote by $n(\theta, \lambda, \rho)$ the needed sample size, which is computed by using the unpooled variance estimate, to detect a risk difference $\delta_*(\theta, \lambda, \rho)$ (see equation (2.3)) at significance level α with $1-\beta$ power.

The expression for $n(\theta, \lambda, \rho)$ is obtained after direct substitution into formula (2.6) and is as follows:

$$n(\theta,\lambda,\rho) = \frac{2\cdot(z_\alpha + z_\beta)^2}{\delta_*(\theta,\lambda,\rho)^2}\cdot\Big(p_*^{(0)}(\theta,\rho)\left(1-p_*^{(0)}(\theta,\rho)\right) + \left(p_*^{(0)}(\theta,\rho)+\delta_*(\theta,\lambda,\rho)\right)\left(1-p_*^{(0)}(\theta,\rho)-\delta_*(\theta,\lambda,\rho)\right)\Big) \tag{2.7}$$

where $p_*^{(0)}(\theta,\rho)$ is given in (2.2). Note that the sample size also relies on the significance level α and the power $1-\beta$, but these are omitted for ease of notation. The corresponding sample size under the pooled estimate can be analogously calculated by using θ, λ and ρ and its expression can be found in the online support material.

2.4.2 Sample size bounds

Assuming that the correlation is the same in the two treatment groups, it follows that the correlation takes values between the lower bound, $B_L(\cdot)$, and the upper bound, $B_U(\cdot)$, which are functions of θ and λ, and are defined as:

$$\begin{aligned} B_L(\theta,\lambda) = \max\Bigg\{ & -\sqrt{\frac{p_1^{(0)}\cdot p_2^{(0)}}{q_1^{(0)}\cdot q_2^{(0)}}}, \quad -\sqrt{\frac{q_1^{(0)}\cdot q_2^{(0)}}{p_1^{(0)}\cdot p_2^{(0)}}}, \\ & -\sqrt{\frac{(p_1^{(0)}+\delta_1)\cdot(p_2^{(0)}+\delta_2)}{(q_1^{(0)}-\delta_1)\cdot(q_2^{(0)}-\delta_2)}}, \quad -\sqrt{\frac{(q_1^{(0)}-\delta_1)\cdot(q_2^{(0)}-\delta_2)}{(p_1^{(0)}+\delta_1)\cdot(p_2^{(0)}+\delta_2)}}\Bigg\} \\ B_U(\theta,\lambda) = \min\Bigg\{ & +\sqrt{\frac{p_1^{(0)}\cdot q_2^{(0)}}{p_2^{(0)}\cdot q_1^{(0)}}}, \quad +\sqrt{\frac{p_2^{(0)}\cdot q_1^{(0)}}{p_1^{(0)}\cdot q_2^{(0)}}}, \\ & +\sqrt{\frac{(p_1^{(0)}+\delta_1)\cdot(q_2^{(0)}-\delta_2)}{(p_2^{(0)}+\delta_2)\cdot(q_1^{(0)}-\delta_1)}}, \quad +\sqrt{\frac{(p_2^{(0)}+\delta_2)\cdot(q_1^{(0)}-\delta_1)}{(p_1^{(0)}+\delta_1)\cdot(q_2^{(0)}-\delta_2)}}\Bigg\} \end{aligned} \tag{2.8}$$

Note that when at least one of the event rates is very close to 0, the lower bound $B_L(\lambda,\theta)$ will also be close to 0 and the plausible correlation values will be always positive. We also notice that, in clinical trials the probabilities of observing the events are often quite low and commonly smaller than 0.5. In this case, the expressions for $B_L(\lambda,\theta)$ and $B_U(\lambda,\theta)$ can be simplified. See the online supplementary material for more details.

Considering such bounds for a given marginal parameters θ and λ, the sample size $n(\theta, \lambda, \rho)$ is an increasing function of the correlation ρ, and it is bounded below and above by $n(\theta, \lambda, B_L(\theta, \lambda))$ and $n(\theta, \lambda, B_U(\theta, \lambda))$, respectively. As a consequence, the more correlated the single endpoints are, the larger will be the necessary sample size for detecting the differences between groups in the composite endpoint. Details for this derivation are provided in Appendix A.2 (see Theorem 1).

2.4.3 Sample size with uncertain correlation value

Since the correlation plays an important role in calculating the sample size, we propose a strategy for deriving the sample size when the parameters that correspond to the composite components are known and the correlation value is not specified in advance.

Prior knowledge about the effect of the treatment being investigated can lead to scientists foreseeing whether the two events of interest, ε_1 and ε_2, are weakly, moderately or strongly correlated. We allow for prior information by splitting the rank of the correlation into three equal-sized intervals, and we consider three correlations categories: weak for the interval whose correlation values are lower; moderate for those intermediate correlation values; and strong for those correlation values that are higher. If any information exists, we will take it into account and will proceed as follows:

(i) *Correlation bounds for each category:*
Considering the categories weak/moderate/strong for the correlation, the plausible correlation values for a given (θ, λ) are in this situation those between the lower and upper values within each category. If the events are weakly correlated, the correlation is between $B_L(\theta, \lambda)$ and $(B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3$; if they are moderately correlated, its value lies between $(B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3$ and $2 \cdot (B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3$; and if they are strongly correlated, it is between $2{\cdot}(B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3$ and $B_U(\theta, \lambda)$. If we cannot place the correlation in any of the above categories, we use the most severe case within its plausible values, then, $B_U(\theta, \lambda)$. (See Table 2.2).

(ii) *Calculate the sample size in each category:*
For the sample size, we advocate using the maximum sample size across all its possible values. That is, $n(\theta, \lambda, (B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3)$, $n(\theta, \lambda, 2\,(B_U(\theta, \lambda) - B_L(\theta, \lambda))\,/3)$, and $n(\theta, \lambda, B_U(\theta, \lambda))$ for weak, moderate

or strong correlations, respectively. Note that since we are assuming the correlation value that maximizes the sample size across its plausible values, we are guaranteeing that the pre-specified power $1-\beta$ is attained.
If the correlation value can not be ascribed to any category, then, we propose a conservative sample size strategy of using the overall possible maximum sample size, that is, $n(\theta, \lambda, B_U(\theta, \lambda))$. Table 2.2 outlines the range of correlations and sample sizes values, together with the proposed sample size for each category.

2.4.4 Sample size accounting for departures from the anticipated event rates

The marginal parameters are often estimated through previous studies or pivotal trials with a limited number of patients and whose patient populations or concomitant drugs could differ from the current ones. Because of that, there is great uncertainty in the values that need to be anticipated for computing the sample size. In this section, we consider that the event rates $p_1^{(0)}$ and $p_2^{(0)}$ have been previously estimated and their corresponding standard errors of the point estimate are provided.

Let $\mathcal{I}_k = \left[\underline{p}_k^{(0)}, \bar{p}_k^{(0)}\right]$ denote a set of plausible values for the true value of $p_k^{(0)}$. For instance, for those previous trials in which we have the standard deviations for the event rates, we can use the set of plausible values for $p_k^{(0)}$ that a 95% confidence interval would yield. We address the issue of sizing a trial for a significance level α and power $1-\beta$ based on the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$, and for fixed effects δ_1 and δ_2 when the correlation value is not known.

We state that, for given δ_1 and δ_2 and at fixed $\rho = r$, the sample size $n(p_1^{(0)}, p_2^{(0)}, \lambda, r)$ (see equation (2.7)) that is needed for power $1-\beta$ at a significance level α, falls into the interval:

$$\mathcal{I}(r, \mathcal{I}_1, \mathcal{I}_2, \lambda) = [\ n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, r), \quad n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, r)\ ] \tag{2.9}$$

This interval is such that it contains the sample size required to attain power $1-\beta$, which is necessary for detecting an effect size equal to $\delta_* = p_*^{(1)} - p_*^{(0)}$ at a significance level α according to the marginal effects δ_1 and δ_2, the correlation r, and the event rates $p_k^{(0)}$ within $\mathcal{I}_k$ $(k = 1, 2)$. Note that the interval $\mathcal{I}(r, \mathcal{I}_1, \mathcal{I}_2, \lambda)$ gives us the plausible sample size values by taking into account the uncertainty of

Table 2.2 Correlation category and its subsequent correlation bounds, $B_L(\cdot)$ and $B_U(\cdot)$ (given in (2.8)) for event rates of the composite components $\theta = (p_1^{(0)}, p_2^{(0)})$, and marginal effect sizes $\lambda = (\delta_1, \delta_2)$. Sample size bounds for each correlation category and proposed sample size strategy calculated by (2.7) according to the margins (θ, λ) and for given significance level α and power $1 - \beta$.

Category	Correlation Bounds	Sample Size Bounds	Sample Size
Weak	$\left[B_L(\theta,\lambda), \frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$	$\left[n(\theta,\lambda,B_L(\theta,\lambda)),\ \ n\left(\theta,\lambda,\frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)\right]$	$n\left(\theta,\lambda,\frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$
Moderate	$\left[\frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}, \frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$	$\left[n\left(\theta,\lambda,\frac{(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right),\ \ n\left(\theta,\lambda,\frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)\right]$	$n\left(\theta,\lambda,\frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right)$
Strong	$\left[\frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}, B_U(\theta,\lambda)\right)$	$\left[n\left(\theta,\lambda,\frac{2(B_U(\theta,\lambda)-B_L(\theta,\lambda))}{3}\right),\ \ n(\theta,\lambda,B_U(\theta,\lambda))\right]$	$n(\theta,\lambda,B_U(\theta,\lambda))$
No prior information	$[B_L(\theta,\lambda), B_U(\theta,\lambda))$	$\left[n(\theta,\lambda,B_L(\theta,\lambda)),\ \ n\left(\theta,\lambda,B_U(\theta,\lambda)\right)\right]$	$n\left(\theta,\lambda,B_U(\theta,\lambda)\right)$

Table 2.3 Correlation category and its subsequent correlation bounds, $\rho_L(\cdot)$ and $\rho_U(\cdot)$ for the intervals of plausible values for event rates, $\mathcal{I}_1 = \left[\underline{p}_1^{(0)}, \bar{p}_1^{(0)}\right]$ and $\mathcal{I}_1 = \left[\underline{p}_2^{(0)}, \bar{p}_2^{(0)}\right]$, and marginal effect sizes $\lambda = (\delta_1, \delta_2)$, and where $\Theta = (\mathcal{I}_1, I_2, \lambda)$ denotes the set of values for the marginal components. Sample size bounds for each correlation category and proposed sample size strategy calculated by (2.7) according to the intervals $\mathcal{I}_1$ and $\mathcal{I}_1$, the marginal effect sizes λ, for given significance level α and power $1 - \beta$.

Category	Correlation Bounds	Sample Size Bounds	Chosen Sample Size
Weak	$\left[\rho_L(\Theta), \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)$	$\left[n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \rho_L(\Theta)),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right)$
Moderate	$\left[\frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \frac{\rho_U(\Theta)-\rho_L(\Theta)}{3}\right),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right)$
Strong	$\left[\frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}, \rho_U(\Theta)\right)$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \frac{2(\rho_U(\Theta)-\rho_L(\Theta))}{3}\right),\ \ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))\right]$	$n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta))$
No prior information	$[\rho_L(\Theta), \rho_U(\Theta))$	$\left[n\left(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \lambda, \rho_L(\Theta)\right),\ \ n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)\right)\right]$	$n\left(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)\right)$

the marginal parameter values, and it provides us the maximum sample size that we would need even though the anticipated event rates are not accurate.

Considering $\Theta = (\mathcal{I}_1, \mathcal{I}_2, \lambda)$ the set of values for the marginal parameters, and denoting by

$$\rho_L(\Theta) = \max_{(\pi_1,\pi_2)\in\mathcal{I}_1\times\mathcal{I}_2} B_L(\pi_1, \pi_2, \lambda)$$
$$\rho_U(\Theta) = \min_{(\pi_1,\pi_2)\in\mathcal{I}_1\times\mathcal{I}_2} B_U(\pi_1, \pi_2, \lambda)$$

the lower and upper bounds of the correlation within the set Θ. Then, for all $(\pi_1, \pi_2) \in \mathcal{I}_1 \times \mathcal{I}_2$, and $\rho \in (\rho_L(\Theta), \rho_U(\Theta))$, we have that:

$$n(\pi_1, \pi_2, \lambda, \rho) \leq U(\Theta) = n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho_U(\Theta)) \tag{2.10}$$

Furthermore, for given $\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda$, the sample size $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$ is an increasing function of the correlation ρ.

The sample size given by $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$ delimits the values that the sample size could have in terms of the correlation accounting for plausible deviations in the anticipated event rates. If there is no prior information on the correlation, we can use $U(\Theta)$ as the needed sample size. If otherwise, we have some prior information on the correlation value, the rationale used in 2.4.3 using correlation categories can be as well applied here to the function $n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, \lambda, \rho)$. Table 2.3 provides the sample size strategy under this circumstance. We lay out the performance of the sample size when varying the event rates in the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ and the subsequent sample size behavior according to the correlation in Propositions 2 and 3 in the supplementary material.

2.4.5 Power performance of the proposed strategies

Given (θ, λ, ρ) and for a fixed sample size N, the power function using the unpooled variance estimate is defined as:

$$\psi(\theta, \lambda, \rho, N) = \Phi\left(\frac{\sqrt{N}\cdot\delta_*(\theta,\lambda,\rho)}{\sqrt{V(\theta,\lambda,\rho)}} - z_\alpha\right) \tag{2.11}$$

where:

$$V(\theta,\lambda,\rho) = p_*^{(0)}\big(\theta,\rho)(1-p_*^{(0)}(\theta,\rho)\big) + \big(p_*^{(0)}(\theta,\rho) + \delta_*(\theta,\lambda,\rho)\big)\big(1-p_*^{(0)}(\theta,\rho) - \delta_*(\theta,\lambda,\rho)\big)$$

and where $\Phi(\cdot)$ denotes the cumulative distribution of the standard normal distribution. The power function for the pooled variance estimator can be found in the online support material.

In what follows, we show that the planned power $1-\beta$ is achieved with any of the previous strategies in Subsections 2.4.3 and 2.4.4.

- If θ and λ are fixed and the correlation value is not known, we have $n(\theta,\lambda,\rho) \leq n\big(\theta,\lambda,B_U(\theta,\lambda)\big)$ and the proposed sample size becomes $N = n\big(\theta,\lambda,B_U(\theta,\lambda)\big)$. The resulting power is then such that:

$$\psi\left(\theta,\lambda,\rho,n\big(\theta,\lambda,B_U(\theta,\lambda)\big)\right) \leq \psi\left(\theta,\lambda,\rho,n(\theta,\lambda,\rho)\right).$$

  The power attained using the upper bound of the correlation is equal to the pre-specified power value $(1-\beta)$ when the correlation ρ is the maximum value within its range, that is, $B_U(\theta,\lambda)$. Otherwise, if the correlation is less than $B_U(\theta,\lambda)$, the power will be always higher than the pre-specified power. Table S1 in the online supplementary material details the power performance when the correlation categories are taken into account.
- If the event rate value $p_k^{(0)}$ is within the interval $\mathcal{I}_k$ for $k=1,2$ and the effect sizes λ are fixed, then $n(p_1^{(0)},p_2^{(0)},\lambda,\rho) \leq n\big(\bar{p}_1^{(0)},\bar{p}_2^{(0)},\lambda,\rho\big)$. If in addition we have no prior information on the correlation value, then since the sample size increases with respect to the correlation, it follows that $n\big(\bar{p}_1^{(0)},\bar{p}_2^{(0)},\lambda,\rho\big) \leq n\big(\bar{p}_1^{(0)},\bar{p}_2^{(0)},\lambda,\rho_U(\Theta)\big)$, and then the proposed sample size turns into $N = n\big(\bar{p}_1^{(0)},\bar{p}_2^{(0)},\lambda,\rho_U(\Theta)\big)$. The corresponding power then satisfies:

$$\psi\left(\theta,\lambda,\rho,n\big(\bar{p}_1^{(0)},\bar{p}_2^{(0)},\lambda,\rho_U(\Theta)\big)\right) \leq \psi\left(\theta,\lambda,\rho,n(p_1^{(0)},p_2^{(0)},\lambda,\rho)\right).$$

  The power attained is equal to the pre-specified power value when the event rates $p_k^{(0)}$ take the upper values $\bar{p}_k^{(0)}$ and the correlation ρ is equal to $\rho_U(\Theta)$. If that is not the case, the power obtained will be larger than the pre-specified $1-\beta$.

2.5 TACTICS-TIMI 18 trial

In managing the syndrome of unstable angina and non-Q-wave acute myocardial infarction, there is controversy over whether using an invasive strategy rather than a conservative strategy offers any advantage. TACTICS-TIMI 18 was a randomized trial that evaluated the efficacy of invasive and conservative treatment

strategies in patients with unstable angina and non-Q-wave AMI treated with tirofiban, heparin, and aspirin (Cannon et al., 2001).

Patients were randomly assigned to either an early invasive strategy or an early conservative strategy. The primary hypothesis of the TACTICS-TIMI 18 trial was that an early invasive strategy would reduce the combined incidence of death, acute myocardial infarction, and rehospitalization for acute coronary syndromes at six months when compared with an early conservative strategy. The primary endpoint was the composite endpoint formed by a combination of incidence of death or non-fatal myocardial infarction (ε_1), and rehospitalization for acute coronary syndrome (ε_2) at six months.

For illustrative purposes, we assume that a trial will be planned for a similar setting and that the results of TACTICS-TIMI 18 are to be used. Since previous studies to TACTICS-TIMI 18 also considered the events death and non-fatal myocardial infarction altogether, we presume that the event rate and effect size on the endpoint ε_1 can be anticipated despite being composed by two events. The estimated values for the frequency of death or non-fatal myocardial infarction (ε_1) in the conservative strategy group was $\hat{p}_1^{(0)} = 0.095$ with a standard deviation of 0.009; whereas the frequency of rehospitalization for acute coronary syndrome (ε_2) was $\hat{p}_2^{(0)} = 0.137$ with a standard deviation of 0.010. Based on the standard deviations of the estimated event rates, we use the 95% confidence intervals as a set of plausible values among which the true values $p_1^{(0)}$, $p_2^{(0)}$ take values, that is, $\mathcal{I}_1 = [0.078, 0.112]$ and $\mathcal{I}_2 = [0.117, 0.157]$. The observed effects on TACTICS-TIMI 18 were $\delta_1 = -0.022$ and $\delta_2 = -0.027$, and we will use these as the expected effects on the new experimental trial.

We consider these parameters to construct the correlation bounds outlined in equation (2.8). The effects δ_1 and δ_2 and the values $\hat{p}_1^{(0)}$ and $\hat{p}_2^{(0)}$ imply that the eligible values for ρ lie in the interval $(-0.10, 0.80)$. Using the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$, the correlation bounds are such that the considered values are plausible for any event rate within $\mathcal{I}_1$ and $\mathcal{I}_2$. This gives us the correlation bounds $(-0.08, 0.77)$. Table 2.5 and Figure 2.1 show the correlation bound according to δ_1 and δ_2 with varying values of the event rates. Observe that the upper bound takes the value 1 when both event rates are equal, and the lower bound tends to 0 when at least one of the event rates becomes smaller.

We illustrate the aspects of calculating power and sample size using the platform CompARE. CompARE calculates the sample size by anticipating the marginal information in terms of either risk difference, relative risk, or odds ratio. In this particular case, we use the statistical test for risk difference under pooled variance in order to ascertain the treatment differences in the composite endpoint at a

Table 2.4 Lower bound, $B_L(\theta, \lambda)$, and upper bound, $B_U(\theta, \lambda)$, for the correlation according to the effect sizes $\delta_1 = -0.022$, $\delta_2 = -0.027$ and for different values of the event rates.

Event rate values	Correlation Bounds
$\hat{p}_1^{(0)} = 0.095,\ \hat{p}_2^{(0)} = 0.137$	$-0.10 \leq \rho \leq 0.80$
$\bar{p}_1^{(0)} = 0.112,\ \bar{p}_2^{(0)} = 0.157$	$-0.12 \leq \rho \leq 0.81$
$\underline{p}_1^{(0)} = 0.078,\ \underline{p}_2^{(0)} = 0.117$	$-0.08 \leq \rho \leq 0.77$

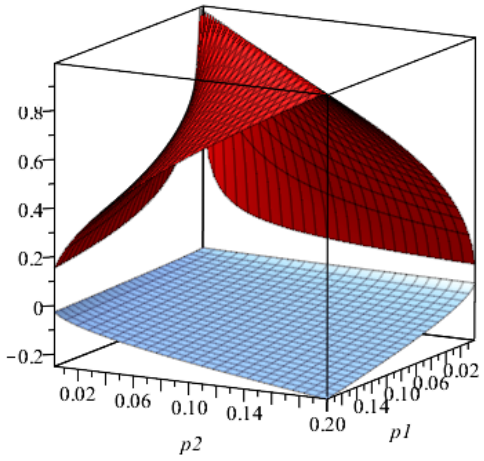


Fig. 2.1 Lower bound (surface in blue) and upper bound (in red) for the correlation according to the effect sizes $\delta_1 = -0.022$, $\delta_2 = -0.027$ and where the marginal event rates take values between 0 and 0.2.

significance level of $\alpha = 0.025$ and target power of $1 - \beta = 0.80$. The results obtained from CompARE are presented in the form of summary tables and plots.

Figure 2.2 (left panel) depicts the performance of the sample size in terms of the correlation for given marginal parameters $\theta = (\hat{p}_1^{(0)}, \hat{p}_2^{(0)})$ and $\lambda = (\delta_1, \delta_2)$; and it illustrates the recommended sample size for each correlation category (weak, moderate, and strong). The solid line represents the sample size as a function of the correlation computed for the anticipated values θ, and the shaded areas represent the region of values, constructed by $\mathcal{I}_1$, $\mathcal{I}_2$, δ_1 and δ_2, within which

interval the sample size falls. Based on $\mathcal{I}_1$ and $\mathcal{I}_2$ the proposed sample size (in dotted lines) is the upper value of the shaded area within the correlation category.

Note that the sample size is highly sensitive to the anticipated parameters. For instance, for $\rho = 0.3$, using $\hat{p}_1^{(0)}$ and $\hat{p}_2^{(0)}$, the required sample size is $n = 3030$. This sample size, however, can differ substantially from that calculated using other reasonable values, such as the upper or lower limits for the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$, which would imply $n = 2511$ and $n = 3540$, respectively.

Figure 2.2 (right panel) describes the statistical power achieved under the proposed method. Assuming that we have correctly anticipated the correlation category, observe that in all cases the achieved power is larger than the planned power, $1 - \beta$. Then, the method guarantees the desired power. If we could correctly anticipate the values of the event rates, then the achieved power would lie between 0.80 and 0.87, in accordance with the plausible correlation values. If we base the sample size calculation on the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$, we will be overestimating the statistical power more than in the previous case, thus obtaining a power between 0.80 and 0.95.

Table 2.5 describes the proposed sample size for each correlation category and reports the possible values for the statistical power, assuming that we have correctly anticipated the correlation category.

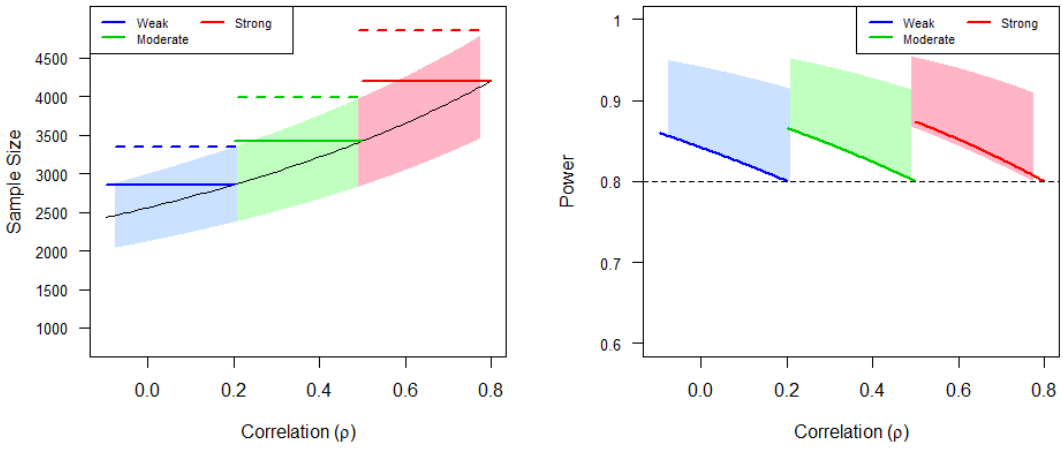
2.6 An extension for risk ratio and odds ratio

In this Section, we show that the risk ratio and odds ratio for the composite endpoint can also be expressed in terms of its margins plus the correlation, and we extend the sample size derivation given in Section 2.4 for evaluating the risk and odds ratio.

2.6.1 Composite effect expressed in terms of the risk ratio or the odds ratio

Let R_k and OR_k denote the risk ratio and odds ratio, respectively, for the k-th event. The risk ratio for the composite endpoint, R_*, is expressed in terms of the risk ratio of its components R_1 and R_2, the event rates under control group, $p_1^{(0)}$ and $p_2^{(0)}$, and the correlation between them, ρ, as follows:

Fig. 2.2 Sample size (left panel) and power (right panel) as a function of the correlation according to the marginal effect sizes $\delta_1 = -0.022$ and $\delta_2 = -0.027$; either based on the point values $\hat{p}_1^{(0)} = 0.095$, $\hat{p}_2^{(0)} = 0.137$ for the event rates (solid line) or based on the interval of plausible values for the event rates $\mathcal{I}_1 = [0.078, 0.112]$ and $\mathcal{I}_2 = [0.117, 0.157]$ (shaded areas). The proposed sample size for each correlation category is highlighted in solid and dotted lines for, respectively, the point values and the interval values for the event rates.



$$\mathrm{R}_* = \frac{p_1^{(0)}\mathrm{R}_1 + p_2^{(0)}\mathrm{R}_2 - p_1^{(0)}p_2^{(0)}\mathrm{R}_1\mathrm{R}_2 - \rho\sqrt{p_1^{(0)}\mathrm{R}_1 p_2^{(0)}\mathrm{R}_2(1 - p_1^{(0)}\mathrm{R}_1)(1 - p_2^{(0)}\mathrm{R}_2)}}{1 - q_1^{(0)}q_2^{(0)} - \rho\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}}} \tag{2.12}$$

Analogously, the odds ratio for the composite endpoint OR_* is defined according to its margins and the correlation is given by:

$$\mathrm{OR}_* = \frac{\left(1+\frac{\mathrm{OR}_1 p_1^{(0)}}{1-p_1^{(0)}}\right)\left(1+\frac{\mathrm{OR}_2 p_2^{(0)}}{1-p_2^{(0)}}\right) - 1 - \rho\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}} \Bigg/ \frac{\left(1+\frac{p_1^{(0)}}{(1-p_1^{(0)})}\right)\cdot\left(1+\frac{p_2^{(0)}}{(1-p_2^{(0)})}\right) - 1 - \rho\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}} \tag{2.13}$$

The derivations of equations (2.12) and (2.13) are postponed to Appendix A.1. By inspection of (2.3), (2.12), and (2.13), we observe that if there is no effect on the

Table 2.5 Recommended sample size for testing differences between the invasive strategy as compared with the conservative strategy. Underlying marginal parameters are as follows: $p_1^{(0)} = 0.095$, $p_2^{(0)} = 0.137$, $\delta_1 = -0.022$, $\delta_2 = -0.027$. Both sample size and power were calculated based on the statistic (2.4) under the pooled variance for a one-sided test at the significance level of $\alpha = 0.025$. The given sample size was calculated to detect the effect on the composite endpoint with the desired overall power of $1 - \beta = 0.80$. For calculating the power of the test, three sample size situations were considered, depending on the strength of the correlation: i) weak correlation; ii) moderate correlation; iii) strong correlation.

Based on point values $p_1^{(0)} = 0.095$, $p_2^{(0)} = 0.137$ for the event rates: Correlation bounds: $B_L(\theta, \lambda) = -0.10$, $B_U(\theta, \lambda) = 0.80$.			
Association strength	**Correlation**	**Sample size**	**Achieved power**
Weak	$-0.10 \leq \rho \leq 0.20$	2860	(0.80, 0.86)
Moderate	$0.20 < \rho \leq 0.50$	3425	(0.80, 0.87)
Strong	$0.50 < \rho \leq 0.80$	4201	(0.80,0.87)
Based on intervals $\mathcal{I}_1 = [0.078, 0.112]$ and $\mathcal{I}_2 = [0.117, 0.157]$ for the event rates: Correlation bounds: $\rho_L(\Theta) = -0.08$, $\rho_U(\Theta) = 0.77$.			
Association strength	**Correlation**	**Sample size**	**Achieved power**
Weak	$-0.08 \leq \rho \leq 0.21$	3355	(0.80, 0.95)
Moderate	$0.21 < \rho \leq 0.49$	3970	(0.80, 0.95)
Strong	$0.49 < \rho \leq 0.77$	4782	(0.80, 0.95)

components, that is, $\delta_1 = \delta_2 = 1$, $\mathrm{R}_1 = \mathrm{R}_2 = 1$ or $\mathrm{OR}_1 = \mathrm{OR}_2 = 1$, then there is no effect on the composite endpoint, $\delta_* = \mathrm{R}_* = \mathrm{OR}_* = 1$. However, the reciprocal does not follow: no effect on the composite endpoint is compatible with some effect on the components. Therefore, it is important to remark, as other authors have warned before (Ferreira-González et al., 2007,b; Tomlinson and Detsky, 2010), that not finding a beneficial effect on composite endpoint is not a guarantee of not having some effect on the components, hence the effect on the composite endpoint cannot be treated as if it were an indicator of some specific effect on its components.

2.6.2 Sample size calculations in terms of risk ratio and odds ratio

The null hypothesis in terms of the risk ratio is stated as $\mathrm{H}_0^* : \log(\mathrm{R}_*) = 0$ and the alternative hypothesis assuming a risk reduction is $\mathrm{H}_1^* : \log(\mathrm{R}_*) < 0$. The statistic

that we use for testing the significance of the relative risk R_* is:

$$\mathrm{Z}_{*,n} = \log(\widehat{\mathrm{R}}_*)/\sqrt{\widehat{Var}(\log(\widehat{\mathrm{R}}_*))}$$

where $\widehat{\mathrm{R}}_* = \widehat{p}_*^{(1)}/\widehat{p}_*^{(0)}$. Under H_0^*, $\mathrm{Z}_{*,n}$ asymptotically follows the standard normal distribution; thus, we will reject H_0^* at the α significance level if $\mathrm{Z}_{*,n} < -z_\alpha$. As in Section 2.3, we estimate the variance $Var(\widehat{\mathrm{R}}_*)$ using the pooled variance by means of $\widehat{Var}_{H_0}(\log(\widehat{\mathrm{R}}_*)) = \frac{2}{n^{(0)}} \cdot \frac{\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}}$ or by using the unpooled variance, $\widehat{Var}_{H_1}(\log(\widehat{\mathrm{R}}_*)) = \frac{1}{n^{(0)}}\left(\frac{1-\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}}{\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}} + \frac{\widehat{q}_*^{(0)}}{\widehat{p}_*^{(0)}}\right)$.

For a given probability under control group $p_*^{(0)}$, and a significance level α, the needed sample size for detecting a risk ratio $\mathit{\Gamma}_* = p_*^{(1)}/p_*^{(0)}$ with power $1-\beta$ is given by:

$$n = 2 \cdot (z_\alpha + z_\beta)^2 \cdot \left(\frac{1-\mathit{\Gamma}_* p_*^{(0)}}{\mathit{\Gamma}_* p_*^{(0)}} + \frac{q_*^{(0)}}{p_*^{(0)}}\right) \Bigg/ \log(\mathit{\Gamma}_*)^2 \tag{2.14}$$

The corresponding sample size when the pooled variance is used can be seen in Table 2.6.

When measuring the effect of treatment with the odds ratio, the null hypothesis $\mathrm{H}_0^* : \log(\mathrm{OR}_*) = 0$ is compared with the alternative hypothesis $\mathrm{H}_1^* : \log(\mathrm{OR}_*) < 0$. To test the above hypotheses we use the statistic:

$$\mathrm{W}_{*,n} = \log(\widehat{\mathrm{OR}}_*)/\sqrt{\widehat{Var}(\log(\widehat{\mathrm{OR}}_*))}$$

where $\widehat{\mathrm{OR}}_* = \frac{\widehat{p}_*^{(1)}/\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}/\widehat{q}_*^{(0)}}$ and where the pooled and unpooled variance estimates are given, respectively, by $\widehat{Var}_{H_0}(\log(\widehat{\mathrm{OR}}_*)) = \frac{8}{n^{(0)}(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)})(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)})}$ and $\widehat{Var}_{H_1}(\log(\widehat{\mathrm{OR}}_*)) = \frac{1}{n^{(0)}}\left(\frac{1}{\widehat{p}_*^{(0)}\widehat{q}_*^{(0)}} + \frac{1}{\widehat{p}_*^{(1)}\widehat{q}_*^{(1)}}\right)$.

Then the needed sample size is calculated using the unpooled variance, for detecting a treatment difference of $\mathrm{OR}_* = \mathit{\Delta}_*$ in order to have power $1-\beta$ at level α for given $p_*^{(0)}$, and it is given by:

$$n = 2 \cdot \left(\frac{z_\alpha + z_\beta}{\log(\mathit{\Delta}_*)}\right)^2 \cdot \left(\frac{(q_*^{(0)} + p_*^{(0)}\mathit{\Delta}_*)^2}{p_*^{(0)} q_*^{(0)} \mathit{\Delta}_*} + \frac{1}{p_*^{(0)} q_*^{(0)}}\right) \tag{2.15}$$

The sample size expression when using the pooled variance can be found in Table 2.6.

Table 2.6 Formulae for sample size determination when comparing two treatments with respect to difference proportions, relative risks or odds ratio contrasts in a balanced design; where n and $n^{(i)}$ denote the total sample size and sample size per group ($i = 0, 1$) needed for testing the effect δ_*, Γ_* or Δ_* for a given event rate within control group $p_*^{(0)}$ at significance level α with $1 - \beta$ power.

	Variance estimator	**Sample Size formula**
Risk difference		
Pooled variance	$\frac{\left(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}\right)\left(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}\right)}{2n^{(0)}}$	$n = 2 \cdot \left(z_\alpha \cdot \sqrt{2\bar{p}_*\bar{q}_*} + z_\beta \cdot \sqrt{p_*^{(0)}q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*)} \right)^2 \Big/ \delta_*^2$
Unpooled variance	$\frac{\left(\widehat{p}_*^{(0)}\widehat{q}_*^{(0)}+\widehat{p}_*^{(1)}\widehat{q}_*^{(1)}\right)}{n^{(0)}}$	$n = 2 \cdot (z_\alpha + z_\beta)^2 \cdot \left(p_*^{(0)}q_*^{(0)} + (p_*^{(0)} + \delta_*)(q_*^{(0)} - \delta_*) \right) \Big/ \delta_*^2$
Risk ratio		
Pooled variance	$\frac{2}{n^{(0)}} \cdot \frac{\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)}}{\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)}}$	$n = 2 \cdot \left(z_\alpha \sqrt{\frac{2\bar{q}_*}{\bar{p}_*}} + z_\beta \sqrt{\frac{1-\Gamma_* p_*^{(0)}}{\Gamma_* p_*^{(0)}} + \frac{q_*^{(0)}}{p_*^{(0)}}} \right)^2 \Big/ \log(\Gamma_*)^2$
Unpooled variance	$\frac{1}{n^{(0)}} \left(\frac{1-\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}}{\widehat{\mathrm{R}}_*\widehat{p}_*^{(0)}} + \frac{\widehat{q}_*^{(0)}}{\widehat{p}_*^{(0)}} \right)$	$n = 2 \cdot (z_\alpha + z_\beta)^2 \cdot \left(\frac{1-\Gamma_* p_*^{(0)}}{\Gamma_* p_*^{(0)}} + \frac{q_*^{(0)}}{p_*^{(0)}} \right) \Big/ \log(\Gamma_*)^2$
Odds ratio		
Pooled variance	$\frac{8}{n^{(0)}(\widehat{p}_*^{(0)}+\widehat{p}_*^{(1)})(\widehat{q}_*^{(0)}+\widehat{q}_*^{(1)})}$	$n = 2 \cdot \left(z_\alpha \sqrt{\frac{2}{\bar{p}_*^{(1)}\bar{q}_*^{(1)}}} + z_\beta \cdot \sqrt{\frac{(q_*^{(0)}+p_*^{(0)}\Delta_*)^2}{p_*^{(0)}q_*^{(0)}\Delta_*} + \frac{1}{p_*^{(0)}q_*^{(0)}}} \right)^2 \Big/ \log(\Delta_*)^2$
Unpooled variance	$\frac{1}{n^{(0)}} \left(\frac{1}{\hat{p}_*^{(0)}\hat{q}_*^{(0)}} + \frac{1}{\hat{p}_*^{(1)}\hat{q}_*^{(1)}} \right)$	$n = 2 \cdot (z_\alpha + z_\beta)^2 \cdot \left(\frac{(q_*^{(0)}+p_*^{(0)}\Delta_*)^2}{p_*^{(0)}q_*^{(0)}\Delta_*} + \frac{1}{p_*^{(0)}q_*^{(0)}} \right) \Big/ \log(\Delta_*)^2$

where: $\bar{p}_* = \frac{p_*^{(0)}+p_*^{(1)}}{2}$ and $\bar{q}_* = \frac{q_*^{(0)}+q_*^{(1)}}{2}$.

2.6.3 Sample size derivation based on its margins

Analogously to Section 2.4 and following the notation in Section 4.4, we obtain the sample size based on the risk ratio as a function of the marginal effects R_1 and R_2, the event rates θ, and the correlation ρ. To do so, we take the event rate and risk ratio of the composite endpoint for their expressions (which are defined according to θ, R_1, R_2 and ρ, see equations (2.2) and (2.12)), and then substitute these into the sample size formula in (2.14). We denote by $n(\theta, R_1, R_2, \rho)$ the needed sample size for evaluating the risk ratio computed for specific values θ, R_1, R_2, and ρ. We will analogously proceed with sample size in terms of the odds ratio using the effects OR_1 and OR_2, then denote by $n(\theta, OR_1, OR_2, \rho)$ the corresponding sample size.

In what follows, we describe the performance of the sample size when the effect is measured by odds ratio or risk ratio. Further details of these properties and their empirical proof are to be found in the web supplementary material.

- For fixed (θ, R_1, R_2) or (θ, OR_1, OR_2), the sample size for testing the effect measured by the risk ratio, $n(\theta, R_1, R_2, \rho)$, and the sample size for testing the odds ratio, $n(\theta, OR_1, OR_2, \rho)$, are increasing functions of the correlation ρ.
- For given R_1 and R_2 at fixed $\rho = r$, the needed sample size $n(p_1^{(0)}, p_2^{(0)}, R_1, R_2, r)$ to have power $1-\beta$ at a significance level α falls into the interval:

$$\mathbf{I}(r, \mathcal{I}_1, \mathcal{I}_2, R_1, R_2) = [\ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, R_1, R_2, r), \quad n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, R_1, R_2, r)\ ] \tag{2.16}$$

  Analogously, for given OR_1 and OR_2, the needed sample size $n(p_1^{(0)}, p_2^{(0)}, OR_1, OR_2, r)$ lies within the interval:

$$\mathbf{I}(r, \mathcal{I}_1, \mathcal{I}_2, OR_1, OR_2) = [\ n(\bar{p}_1^{(0)}, \bar{p}_2^{(0)}, OR_1, OR_2, r), \quad n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, OR_1, OR_2, r)\ ]$$

- For all $(\pi_1, \pi_2) \in \mathcal{I}_1 \times \mathcal{I}_2$ and $\rho \in (\rho_L(\Theta), \rho_U(\Theta))$, it follows that:

$$n(\pi_1, \pi_2, R_1, R_2, \rho) \leq \mathcal{U}_R(\Theta) = n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, R_1, R_2, \rho_U(\Theta))$$
$$n(\pi_1, \pi_2, OR_1, OR_2, \rho) \leq \mathcal{U}_{OR}(\Theta) = n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, OR_1, OR_2, \rho_U(\Theta))$$

  Furthermore, for given $(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, R_1, R_2)$ or $(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, OR_1, OR_2)$, the sample size functions $n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, R_1, R_2, \rho)$ and $n(\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, OR_1, OR_2, \rho)$ increase with respect to the correlation ρ.

Note that, unlike when using risk differences, the sample size has its maximum value when both event rates take their lower interval values $\underline{p}_1^{(0)}, \underline{p}_2^{(0)}$ (see equations (2.9) and (2.16)).

Also note that if the marginal parameters $(\theta, \mathrm{R}_1, \mathrm{R}_2)$ or $(\theta, \mathrm{OR}_1, \mathrm{OR}_2)$ are anticipated and the correlation is not known, the sample size strategy described in Section 2.4.3 can be extended to the risk and odds ratio and analogously applied. For fixed effects $(\mathrm{R}_1, \mathrm{R}_2)$ or $(\mathrm{OR}_1, \mathrm{OR}_2)$, and given intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ for the event rates, we can follow the same reasoning as for risk differences in Section 2.4.4, and use $\mathcal{U}_R(\Theta)$ (analogously $\mathcal{U}_{OR}(\Theta)$) to calculate the required sample size that guarantees the planned power while accounting for the unknown correlation value and uncertainty of the marginal parameter values.

2.7 A simulation study

We conduct a simulation study to evaluate the strategies proposed in Section 2.4 for computing the sample size.

2.7.1 Design

We simulate a two-arm trial with a composite primary endpoint composed of two events, ε_1 and ε_2, according to the following values (which are all summarized in Table 2.7): the marginal probabilities of observing ε_k $(k = 1, 2)$ in the control group $\theta = (p_1^{(0)}, p_2^{(0)})$ take values between 0.01 and 0.2, and they are without loss of generality such that $p_1^{(0)} < p_2^{(0)}$; the risk ratios $\lambda = (\mathrm{R}_1, \mathrm{R}_2)$ are specified for beneficial effects and vary from 0.6 to 0.8; the true correlation between ε_1 and ε_2 is assumed to be common for both groups, and it covers the positive range between 0 and 1. The possible combinations add up to a total of 421 different scenarios which take into account that for given (θ, λ). Simulations are performed only for those ρ_{true} between $B_L(\theta, \lambda)$ and $B_U(\theta, \lambda)$ (see (2.8)).

For each one of these 421 scenarios specified by $(\theta,\ \lambda,\ \rho_{true})$, we compute the required sample size $n(\theta, \lambda, \rho(\theta, \lambda))$ for a one-sided test with power $1-\beta = 0.80$ at the significance level $\alpha = 0.025$, which is done by following one of the six different formulations that are derived in Section 2.3 and Section 2.6 and, additionally, are all summarized in Table 2.6.

We distinguish 4 different situations according to the value we assume in $\rho(\theta, \lambda)$ to calculate $n(\theta, \lambda, \rho(\theta, \lambda))$:

(1) For the weak correlation category, use $\rho(\theta, \lambda) = B_U(\theta, \lambda)/3$
(2) For the moderate correlation category, use $\rho(\theta, \lambda) = 2B_U(\theta, \lambda)/3$
(3) For the strong correlation category, use $\rho(\theta, \lambda) = B_U(\theta, \lambda)$
(4) For guessing the true correlation, use $\rho(\theta, \lambda) = \rho_{true}$.

Given one scenario specified by (θ, $\lambda = (R_1, R_2)$, ρ_{true}), we evaluate the type I error first by calculating n based on $(\theta, \lambda = (R_1, R_2), \rho(\theta, \lambda))$ and simulating 100000 trials using $(\theta, \lambda = (1, 1), \rho_{true})$. To check the power, we start by calculating n as above, based on $(\theta, \lambda = (R_1, R_2), \rho(\theta, \lambda))$, and then we simulate 100000 trials using $(\theta, \lambda = (R_1, R_2), \rho_{true})$. Altogether, we have to analyze a total of 3368 scenarios.

The above steps have to be reproduced six times according to the different sample size formulae used to compute $n(\theta, \lambda, \rho(\theta, \lambda))$, that is, by stating the effect in terms of the difference in proportions, the risk ratios or the odds ratio, and using both the pooled and the unpooled estimates of the variance. We have performed all computations using the `R` software tool (Version 0.98.1087), and the time required to perform the considered simulations was 55.58h.

Table 2.7 Simulation scenarios: Values of marginal event rates in the control group: $\theta = (p_1^{(0)}, p_2^{(0)})$; treatment effects in terms of the risk ratio: $\lambda = (\mathrm{R}_1, \mathrm{R}_2)$; and correlation ρ_{true} between components. Note that not all the combinations are feasible because the correlation is between $B_L(\theta, \lambda)$ and $B_U(\theta, \lambda)$.

Parameter	**Values**
$p_1^{(0)}$	$0.01, 0.05, 0.10$
$p_2^{(0)}$	$0.01, 0.05, 0.10, 0.15, 0.20$
ρ_{true}	$0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$
Effects used to evaluate the power:	
$\mathrm{R}_1, \mathrm{R}_2$	$0.6, 0.7, 0.8$
Effect used to evaluate the type I error:	
$\mathrm{R}_1 = \mathrm{R}_2$	1

2.7.2 Power analysis of the proposed strategies for computing sample size

Let $n_{l,m}$ be the required sample size calculated using the formulae described in Table 2.6 where $l = p, u$, indicating whether the pooled or unpooled variance has been used; and $m = D, R, OR$, indicating the effect measure that has been tested. In other words, for the difference in proportions, $m = D$; for relative risk, $m = R$; and for odds ratio, $m = OR$. Let $\Psi_{l,m}$ denote the empirical power when the total number of participants is $n_{l,m}$ $(l = p, u;\ \ m = D, R, OR)$.

Whenever the correlation we are using to compute the sample size coincides with the one we have used to run the simulations ($\rho(\theta, \lambda) = \rho_{true}$), the empirical powers are always achieved whether we are using the pooled, $\Psi_{p,m}$, or unpooled, $\Psi_{u,m}$, estimator of the variance. Nevertheless, when testing the difference in proportions, the achieved powers do not substantially differ ($\Psi_{u,D} \cong \Psi_{p,D}$); when testing the treatment differences in terms of the risk ratio or the odds ratio, the power achieved if the unpooled variance estimator is used is slightly larger than the power achieved with the pooled estimator, $\Psi_{u,m} \geq \Psi_{p,m}$, $m = R, OR$ (see Table S3 in supplementary material for a comparison of the two approaches). The results presented herein refer to the unpooled variance estimator. The corresponding results for the pooled variance are summarized in the supplementary material (Table S4 and Figure S1).

When $\rho(\theta, \lambda) \neq \rho_{true}$, we distinguish two types of misspecification. Misspecification type I, ρ_{true} and $\rho(\theta, \lambda)$ pertain to the same correlation category; and Misspecification type II, ρ_{true} and $\rho(\theta, \lambda)$ do not belong to the same category. Table 2.8 describes the empirical power in these two cases, which account for the correlation category for the three effect measures that we could use to test the difference between groups. If Misspecification I occurs, the pre-specified power is achieved and might exceed 7%.

For misspecification II, there are two possible scenarios. The first is for those cases where the correlation $\rho(\theta, \lambda)$ is assumed in a stronger correlation category than the one that ρ_{true} belongs to, for instance, if $\rho(\theta, \lambda)$ is assumed to be strong and ρ_{true} is moderate. Under this scenario, $\rho(\theta, \lambda) > \rho_{true}$, and then the planned power is always achieved. The second scenario is when the $\rho(\theta, \lambda)$ is assumed to be in a weaker correlation category than the one that ρ_{true} lies in. For instance, when $\rho(\theta, \lambda)$ is assumed weak and ρ_{true} is moderate. In those cases where $\rho(\theta, \lambda) < \rho_{true}$, the trial will be underpowered.

The empirical power in terms of the difference between the assumed and true correlations is illustrated in Figure 2.3. Observe that when the assumed corre-

lation is greater than the true correlation, that is, $\rho(\theta, \lambda) > \rho_{true}$, the empirical power is equal to or greater than the pre-specified power. Note that in all cases under the strong correlation category we have $\rho_{true} \leq \rho(\theta, \lambda)$, the pre-specified power is assured even though we failed to anticipate the category. Also note that there are no differences in the achieved power, nor are there any in the method's performance in terms of the measure we are using to evaluate the effect.

Table 2.8 Median empirical power, given the sample size (under the unpooled variance), depending on the misspecification error and the assumed correlation. Values in parentheses indicate the maximum and minimum of the empirical power.

Assumption	**Misspecification I: Correlation within the category**	**Misspecification II: Correlation outside the category**
Risk Difference		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.80)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.91)
Strong	0.82 (0.80, 0.87)	0.87 (0.81, 0.95)
Risk Ratio		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.81)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.90)
Strong	0.82 (0.80, 0.87)	0.88 (0.81, 0.95)
Odds Ratio		
Weak	0.82 (0.80, 0.86)	0.78 (0.67, 0.81)
Moderate	0.82 (0.80, 0.87)	0.82 (0.74, 0.91)
Strong	0.82 (0.80, 0.87)	0.87 (0.81, 0.95)

2.7.3 Type I error analysis of the proposed strategies for computing sample size

The empirical results in the simulation study show that the type I error is not affected by the misspecification of the correlation. Nonetheless, the empirical type I error under the pooled estimator may be slightly superior to significance level 0.025, especially when the treatment is tested in terms of risk ratio and odds ratio (see Figure S2 in the online support material).

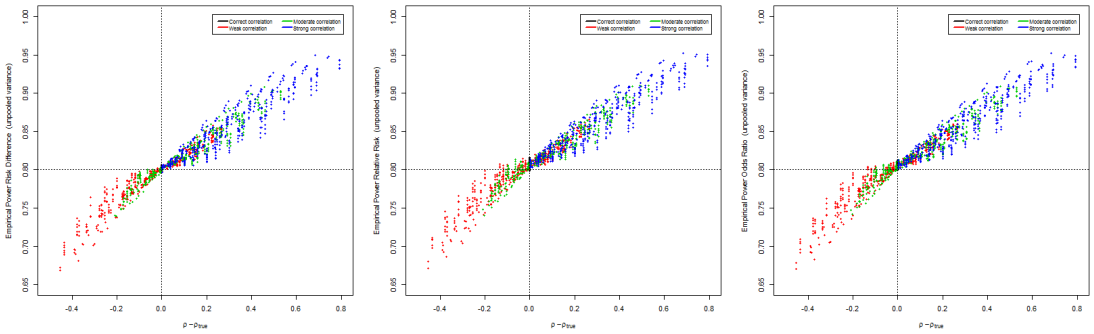


Fig. 2.3 Scatterplot showing the relationship between empirical power versus the difference between the assumed and true correlations for each of the sample size formulas (under unpooled variance) that were used in the simulation study in section 2.7.

2.8 Supplementary material

Additional information of this work may be found online in the Supporting material, which is available at:

`https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8092`

The supplementary material contains: the sample size and power formulae for the pooled variance estimator; a short discussion on the consequences on the assumption $\rho^{(0)} = \rho^{(1)}$; the derivation of the correlation bounds given in (2.8) (see Section 2.4); an empirical demonstration of the sample size behavior when varying the event rates in the intervals I_1 and I_2 outlined in Section 2.4.4; and a study of the sample size performance when using I_1 and I_2.

The source code to reproduce the results may be found online in the GitHub repository:

`https://github.com/CompARE-Composite/Functions`

2.9 Further work

This section has two main parts which include some work that has not been published. First, we introduce different association measures for binary endpoints –aside from Pearson's correlation–, describe some of their characteristics and define the probability of the composite endpoint by means of them. Second, we

outline some properties of the composite effect aiming at shedding some light on the interpretation of the composite effect.

2.9.1 On the association between binary endpoints

Pearson's correlation is the most common measure to quantify the degree of association between binary endpoints, there are, however, more intuitive alternative measures to define the association between two binary outcomes. In this section, we present the relative overlap and the conditional probability as different ways to measure the association between two binary endpoints, and we rewrite $p_*^{(i)}$ in terms of each of them.

Relative overlap

The relative overlap in the i-th group of treatment is defined as the conditional probability of observing the two marginal events knowing that at least one of these events has occurred (Marsal et al., 2015). This measure is evaluated as the ratio between the probability of the intersection and the probability of the composite endpoint, as follows:

$$RO^{(i)} = \frac{p_{\cap}^{(i)}}{p_1^{(i)} + p_2^{(i)} - p_{\cap}^{(i)}} \tag{2.17}$$

Note that this measure quantifies the ratio of the intersection versus the union of having these two events. The relative overlap takes values between 0 and 1 and is bounded by:

$$RO^{(i)} \in \left[\max\{0, p_1^{(i)} + p_2^{(i)} - 1\}, \quad \min\left\{ \frac{p_1^{(i)}}{p_2^{(i)}}, \frac{p_2^{(i)}}{p_1^{(i)}} \right\} \right] \subseteq [0, 1] \tag{2.18}$$

Clinical studies that include probabilities of observing the events smaller than 0.5 are common in many disease areas such as oncology, and cardiovascular disease. We note that for $p_1^{(i)} \leq p_2^{(i)} \leq 0.5$, the bounds (2.18) can be simplified to:

$$RO^{(i)} \in \left[0, \quad \frac{p_1^{(i)}}{p_2^{(i)}} \right]$$

In the following proposition, we establish the relationship between the relative overlap and the correlation and underline some properties.

Proposition 2.1. *Let $RO^{(i)}$ and $\rho^{(i)}$ be the relative overlap and the correlation in the i-th group, respectively. Then, the following is true:*

(i) The relative overlap, $RO^{(i)}$, can be expressed by means of the event rates, $p_1^{(i)}$ and $p_2^{(i)}$, and the correlation, $\rho^{(i)}$, as follows:

$$RO^{(i)} = \frac{\rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}} + p_1^{(i)}p_2^{(i)}}{1 - q_1^{(i)}q_2^{(i)} - \rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}}}$$

(ii) If the correlation is equal zero, the relative overlap is not, even when there is independence between the events. That is:

$$\rho^{(i)} = 0 \nRightarrow RO^{(i)} = 0$$

(iii) The relative overlap is zero if, and only if, the intersection probability is zero and, hence, when the correlation is negative. That is:

$$RO^{(i)} = 0 \Leftrightarrow p_\cap^{(i)} = 0 \quad \textit{and} \quad p_\cap^{(i)} = 0 \Rightarrow \rho^{(i)} < 0$$

Proof (Proof of Proposition 2.1). Consider $\rho^{(i)}$, defined as:

$$\rho^{(i)} = \frac{p_\cap^{(i)} - p_1^{(i)}p_2^{(i)}}{\sqrt{p_1^{(i)}q_1^{(i)}p_2^{(i)}q_2^{(i)}}} \tag{2.19}$$

The proof of (i) is straightforward by solving (2.19) by $p_\cap^{(i)}$, and then plugging this back into the relative overlap definition given in (2.17). The proof of (ii) and (iii) follows directly from the relationship stated in (i).

In the next result, we state how the probability of the composite endpoint can be rewritten in terms of the marginal parameters and the relative overlap.

Proposition 2.2. *The probability of having the composite event in terms of the relative overlap and the event rates of the marginal parameters is given by:*

$$p_*^{(i)} = \frac{p_1^{(i)} + p_2^{(i)}}{1 + RO^{(i)}}$$

Proof. Consider first the definition of $RO^{(i)}$ given in (2.17). Noticing that $p_*^{(i)} = p_1^{(i)} + p_2^{(i)} - p_\cap^{(i)}$, we have:

$$RO^{(i)} = \frac{p_1^{(i)} + p_2^{(i)} - p_*^{(i)}}{p_*^{(i)}} \tag{2.20}$$

Then, the expression given in the proposition is obtained by solving the previous equation for $p_*^{(i)}$.

Note that the probability of the composite endpoint is inversely proportional to the relative overlap.

Conditional probability

The conditional probability of observing the two marginal events is the probability of an event occurring given that another event has occurred. Let $P_{X_1|X_2}^{(i)}$ denote the conditional probability of X_1 given X_2 in group i, defined as follows:

$$P_{X_1|X_2}^{(i)} = P(X_{ij1} = 1 | X_{ij2} = 1) = \frac{p_\cap^{(i)}}{p_2^{(i)}} \tag{2.21}$$

Note that this measure quantifies the ratio of the intersection probability over the probability of having had the event ε_2. Also note that the conditional probability is that is not symmetrical with respect to the role of X_1 and X_2 plays on it. Indeed, we have that:

$$P_{X_1|X_2}^{(i)} = P_{X_2|X_1}^{(i)} \cdot \frac{p_1^{(i)}}{p_2^{(i)}}$$

where $P_{X_2|X_1}^{(i)}$ is the conditional probability of X_2 given X_1.

The conditional probability takes values between 0 and 1 and is parametrically bounded:

$$P_{X_1|X_2}^{(i)} \in \left[p_1^{(i)}, \min\left\{ \frac{p_1^{(i)}}{p_2^{(i)}}, 1 \right\} \right] \tag{2.22}$$

Note that, if $p_1^{(i)} < p_2^{(i)}$, then the bounds are simplified to $\left[p_1^{(i)}, \frac{p_1^{(i)}}{p_2^{(i)}}\right]$; otherwise, $\left[p_1^{(i)}, 1\right]$.

We state the relationship between the conditional probability and the correlation in the following proposition.

Proposition 2.3. *Let $P^{(i)}_{X_1|X_2}$ be the conditional probability of X_1 given X_2 and $\rho^{(i)}$ be the correlation in the i-th group. Then, the following is true:*

(i) The relationship between the conditional probability and the correlation is given by:

$$P^{(i)}_{X_1|X_2} = \frac{p^{(i)}_1 \cdot p^{(i)}_2 + \rho^{(i)}\sqrt{p^{(i)}_1 p^{(i)}_2 q^{(i)}_1 q^{(i)}_2}}{p^{(i)}_2} = p^{(i)}_1 + \rho^{(i)} \cdot \frac{\sqrt{p^{(i)}_1 p^{(i)}_2 q^{(i)}_1 q^{(i)}_2}}{p^{(i)}_2}$$

(ii) If the correlation is zero, the conditional probability is not. Indeed,

$$\rho^{(i)} = 0 \Rightarrow P^{(i)}_{X_1|X_2} = p_1$$

(iii) If $\rho = \rho_{\max}$, $P^{(i)}_{X_1|X_2}$ depends on the marginal probabilities: if $p^{(i)}_2 > p^{(i)}_1$, then $P^{(i)}_{X_1|X_2} = \frac{p^{(i)}_1}{p^{(i)}_2}$; otherwise $P^{(i)}_{X_1|X_2} = 1$.

(iv) The conditional probability is zero if, and only if, the intersection probability is zero and, hence, when the correlation is negative.

$$P^{(i)}_{X_1|X_2} = 0 \Leftrightarrow p^{(i)}_{\cap} = 0 \quad \textit{and} \quad p^{(i)}_{\cap} = 0 \Rightarrow \rho^{(i)} < 0$$

Proof. Let us consider $\rho^{(i)}$ defined in (2.19). The proof of (i) follows directly by solving (2.19) by $p^{(i)}_{\cap}$, and then plugging this back into the conditional probability definition given in (2.21). The proofs of (ii), (iii) and (iv) come from the relationship stated in (i).

Next, we show that the probability of the composite endpoint can be expressed as well in terms of the conditional probabilities and the probabilities of the composite components.

Proposition 2.4. *The probability of having the composite event in terms of the conditional probability and the probability of its components is given by:*

$$p^{(i)}_* = p^{(i)}_1 + p^{(i)}_2 - P^{(i)}_{X_1|X_2} \cdot p^{(i)}_2$$

Proof (Proof of Proposition 2.4). The proof follows by noticing that:

$$p^{(i)}_* = p^{(i)}_1 + p^{(i)}_2 - p^{(i)}_{\cap} = p^{(i)}_1 + p^{(i)}_2 - P^{(i)}_{X_1|X_2} \cdot p^{(i)}_2 \tag{2.23}$$

Note that the composite probability decreases with respect to $P^{(i)}_{X_1|X_2}$. Also notice that taking into account (2.22), we have:

$$p^{(i)}_* \in \left[\max\{p^{(i)}_1, p^{(i)}_2\},\ \ p^{(i)}_1 + p^{(i)}_2 - p^{(i)}_1 \cdot p^{(i)}_2\right]$$

Table 2.9 Association measures between binary endpoints.

Definition	**Association bounds**	**Composite probability**
$\rho^{(i)} = \frac{p_{\cap}^{(i)} - p_1^{(i)} p_2^{(i)}}{\sqrt{p_1^{(i)} p_2^{(i)} q_1^{(i)} q_2^{(i)}}}$	$\left[\max\left\{ -\sqrt{\frac{p_1^{(i)} \cdot p_2^{(i)}}{q_1^{(i)} \cdot q_2^{(i)}}}, \; -\sqrt{\frac{q_1^{(i)} \cdot q_2^{(i)}}{p_1^{(i)} \cdot p_2^{(i)}}} \right\}, \; \min\left\{ \sqrt{\frac{p_1^{(i)} \cdot q_2^{(i)}}{p_2^{(i)} \cdot q_1^{(i)}}}, \; \sqrt{\frac{p_2^{(i)} \cdot q_1^{(i)}}{p_1^{(i)} \cdot q_2^{(i)}}} \right\} \right]$	$p_*^{(i)} = 1 - q_1^{(i)} q_2^{(i)} - \rho^{(i)} \sqrt{p_1^{(i)} p_2^{(i)} q_1^{(i)} q_2^{(i)}}$
$RO^{(i)} = \frac{p_{\cap}^{(i)}}{p_1^{(i)} + p_2^{(i)} - p_{\cap}^{(i)}}$	$\left[\max\{0, p_1^{(i)} + p_2^{(i)} - 1\}, \; \min\left\{ \frac{p_1^{(i)}}{p_2^{(i)}}, \frac{p_2^{(i)}}{p_1^{(i)}} \right\} \right]$	$p_*^{(i)} = \frac{p_1^{(i)} + p_2^{(i)}}{1 + RO^{(i)}}$
$P_{X_1 \mid X_2}^{(i)} = \frac{p_{\cap}^{(i)}}{p_2^{(i)}}$	$\left[p_1^{(i)}, \min\left\{ \frac{p_1^{(i)}}{p_2^{(i)}}, 1 \right\} \right]$	$p_*^{(i)} = p_1^{(i)} + p_2^{(i)} - P_{X_1 \mid X_2}^{(i)} p_2^{(i)}$

2.9.2 On the magnitude of the composite effect

By inspection of the expression of the composite endpoint (see equations 2.3, 2.12, and 2.13), we observe that if there is no effect on the components, then there is no effect on the composite endpoint. However, not finding a beneficial effect on composite endpoint is not a guarantee of not having some effect on the components, hence the effect on the composite endpoint cannot be treated as if it was an indicator of which is the effect on its components.

Under the particular case of independence between the components, the effect on the composite endpoint in terms of the odds ratio takes values between the margins effects, that is, assuming $\mathrm{OR}_1 \leq \mathrm{OR}_2$, then $\mathrm{OR}_1 \leq \mathrm{OR}_* \leq \mathrm{OR}_2$, and analogously for the odds ratio (See Theorem 2.1). But if the components are correlated, this property is not assured, then the composite effect could be greater (or smaller) than both marginal effects, i.e., $\mathrm{OR}_* < \mathrm{OR}_1 \leq \mathrm{OR}_2$ (or $\mathrm{OR}_* > \mathrm{OR}_1 \geq \mathrm{OR}_2$).

In many trials the composite components represent relatively rare events (FDA Guidance, 2017), since studying each component separately would require unmanageable large sample sizes. In those cases when there is one event –let's say event ε_1– whose event rate is low and smaller than the second event ($p_1^{(0)} < 0.1$ and $p_1^{(0)} < p_2^{(0)}$), the effect on the composite endpoint is approximately equal to the treatment effect for the second event, that is, $\mathrm{OR}_* \cong \mathrm{OR}_2$ (see Theorem 2.2). Nevertheless, when there are different frequencies and effect sizes on the composite components, this effect should be approached carefully because it may strongly dependent on the correlation.

Theorem 2.1. *If the independence assumption holds between the composite components, the correlation is the same in the two groups, that is,* $\rho^{(0)} = \rho^{(1)} = 0$*, and if* $\mathrm{OR}_1, \mathrm{OR}_2 < 1$ *(beneficial effect), then:*

$$\mathrm{OR}_* \in [\min(\mathrm{OR}_1, \mathrm{OR}_2), \max(\mathrm{OR}_1, \mathrm{OR}_2)] \tag{2.24}$$

Proof (Proof of Theorem 2.1). Denote by $\mathrm{O}_k^{(i)}$ the odds for each endpoint in the i-th group, that is, $\mathrm{O}_k^{(i)} = p_k^{(i)}/q_k^{(i)}$. Then, the expression (A.4) become in this case:

$$\mathrm{OR}_* = \frac{\mathrm{OR}_1\mathrm{O}_1^{(0)} + \mathrm{OR}_2\mathrm{O}_2^{(0)} + \mathrm{OR}_1\mathrm{OR}_2\mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}}{\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}}$$

Let M denote $\max(\mathrm{OR}_1, \mathrm{OR}_2)$, then $\mathrm{OR}_k\mathrm{O}_k^{(0)} \leq M\mathrm{O}_k^{(0)}$, and it follows:

$$\begin{aligned}\mathrm{OR}_* &\leq \frac{M\mathrm{O}_1^{(0)} + M\mathrm{O}_2^{(0)} + M^2\mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}}{\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}} = \frac{M(\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + M\mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)})}{\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}} \\ &\leq \frac{M(\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)})}{\mathrm{O}_1^{(0)} + \mathrm{O}_2^{(0)} + \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}} = M = \max(\mathrm{OR}_1, \mathrm{OR}_2)\end{aligned}$$

where the last inequality follows from $M\mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)} < \mathrm{O}_1^{(0)}\mathrm{O}_2^{(0)}$ whenever $\mathrm{OR}_k < 1$. If follows analogously that $\mathrm{OR}_* \geq \min(\mathrm{OR}_1, \mathrm{OR}_2)$.

Theorem 2.2. *Let $p_1^{(0)}$ and $p_2^{(0)}$ be the probabilities of observing the individual components of the composite in the control group, and let OR_1 and OR_2 be the odds ratio for the individual components. Denoting by $\mathrm{OR}_*(p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho)$ the odds ratio function described in* (A.4) *in terms of the event rates, odds ratio and correlation, the odds ratio for the composite endpoint tends to OR_1, as $p_2^{(0)}$ tends to 0.*

Proof (Proof of Theorem 2.2). The proof is straightforward by noting that:

$$\lim_{p_2^{(0)} \to 0} \mathrm{OR}_*(p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho) = \mathrm{OR}_1 \tag{2.25}$$

As a consequence, for every $\epsilon > 0$, there exists a $\delta > 0$ such that $\left|\mathrm{OR}_*(p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho) - \mathrm{OR}_1\right| < \epsilon$, for all $p_2^{(0)} < \delta$.

2.10 Discussion

Composite endpoints are increasingly used as primary endpoints to achieve greater incidence rates of observing the primary event, larger effect sizes and, hopefully, higher statistical power while avoiding multiplicity adjustment. Even so, their use creates challenges in both the design and interpretation of the studies.

It is well known that sample size determination plays a key role in trial design. We have shown that calculating the sample size for composite binary endpoints needs more than the anticipated effect size and event rates of the composite components; it also needs the correlation between them. Sizing clinical trials in which composite endpoints are involved often implies facing the challenge of dealing with the unknown value of the correlation. We have assessed how much the correlation impacts the sample size and, consequently, the attained power. Our conclusion is that the sample size strongly depends on the correlation and that the more

correlation between the components are, the more sample size is needed. Motivated by this concern, we have proposed some strategies for deriving the sample size when the correlation is not specified. The strategy, based on the stratification of the correlation into different categories, assures the pre-specified power even without previous knowledge on the correlation. In addition, if at least we could anticipate the category where the correlation falls into, the achieved power would slightly surpass the planned power (see Table 2.8). In those cases where not even the correlation category can be anticipated, the interval of plausible values for the sample size might be too wide and the proposed strategy might be extremely conservative. Further research is needed in such cases to obtain more accurate power.

We have illustrated our proposal using the platform CompARE. CompARE is an open-source and completely free web platform that can be used as a tool for clinicians, medical researchers and statisticians to compute the sample size according to the procedure proposed in this work. We will present CompARE and describe its features in Chapter 5.

Throughout this work, we have assumed that we are in the planning stage of a randomized clinical trial whose aim is to test the efficacy of a new treatment by comparing its performance with others that have already been approved. These trials are usually known to have much larger sample sizes. For this reason, we have restricted this work to sample size calculation based upon asymptotic approximations of the normal distribution. In previous trial phases devoted to obtain the optimal dose level or to study the toxicity of the new drug, the sample size is not as large as in efficacy trials. In those cases, it could be more appropriate to base the sample size calculations on an exact test. Unlike the tests based on asymptotic distributions, the power function of an exact test usually does not have an explicit form, and the sample size is obtained numerically by greedy search over the sample space. In practice, the applicability of such methods can come across difficulties because intensive computing is required (Chow et al., 2008). There is controversy over whether or not to use exact tests, since when the sample size is not large enough, the asymptotic test may not preserve the test size, whereas exact tests could be conservative (Fagerland et al, 2015; Crans and Shuster, 2008; Andrés and Tejedor, 2009).

The sample size calculation in this work has been derived using the same correlations for both groups. Although this assumption is very often being used (Sugimoto et al., 2017; Asakura et al., 2017; Ando et al., 2015), it remains to be studied how plausible is in practice. We are working on an extension of our methods to account for different group correlations. Moreover, we are currently

studying and implementing in CompARE other association measures for characterizing the strength of dependence between pairs of binary endpoints, such as the ones presented in Section 2.9, which in practice might be easier to anticipate.

Interpreting the results of a trial with a primary composite endpoint is particularly challenging. Composite endpoints comprise the information of its components and capture a more complex picture of the intervention's efficacy, however, they might oversimplify the evidence by looking only at the composite effect (Pocock et al., 2015). A proper study of the contribution from each separate component should be conducted to ensure a clear understanding of the results. What is more, composite endpoints are in many cases formed by a set of endpoints among whom the clinical relevance highly differs. This could lead to misleading results about whether the treatment benefits only the less important endpoints. Moreover, as shown in Sections 2.6 and 2.9, the effect for the composite does not necessarily reflect the effects for the components. As we will see in Chapter 5, CompARE computes the effect on the composite endpoint and gets constructive numerical and graphical results in order to investigate the role that each component plays.

Different strategies such as the win ratio (Pocock et al., 2012; Luo et al., 2015) and the weighted combined approach (Rauch et al., 2017) have been developed to take into consideration the order of clinical priorities for the composite components when analyzing composite endpoints. Extending this work to more than two components and by incorporating weights remains open for future research.

Chapter 3

Endpoint selection on composite binary endpoints

The main content of this chapter has been published in:

Selection of composite binary endpoints in clinical trials.
Bofill Roig, M., and Gómez Melis, G.
Biometrical Journal. Volume 60, Issue 2, March 2018, Pages 246-261.
DOI: 10.1002/bimj.201600229.

Sections 3.1 to 3.6 reproduce almost identically the corresponding work in the paper. However, some modifications have been done for coherence and cohesion of the thesis. Section 3.7 describes further work that has not been published. The discussion corresponds mostly to the one in the paper, but we have slightly changed some paragraphs to update and encompass the further work we have made. The methodology presented in this chapter has been implemented in CompARE, but we postpone the software's explanation to Chapter 5. Additional tables and figures may be found in Appendix B. The code to reproduce the results of this work may be found online at: `https://github.com/MartaBofillRoig/CompARE`.

3.1 Introduction

Nowadays, randomized clinical trials guide the advance of medical knowledge. They are the most well-grounded procedure for evaluating the applicability of clinical research and also comparing the safety and effectiveness of a new intervention against the standard of care. The protocol formalizes the medical question and specifies the design of the trial. One key decision that has to be defined is the choice of the primary endpoint which measures the clinical evidence by quantifying the treatment effect. Sample size requirement, analysis and conclusions on efficacy are based on the primary endpoint.

Clinical trials often take into account two or more efficacy endpoints. If we use multiple co-primary endpoints, we could capture different attributes. Moreover, multiple co-primary endpoint might provide a better explanation about how the disease behaves under treatment and an improvement of the evaluation of whether there are the differences in the efficacy between different interventions. However, the use of multiple endpoints entails challenges in analyses and planning. On the one hand, we need a multiplicity adjustment for avoiding an inflation of the overall type I error (Pocock et al., 1987). On the other hand, multiple co-primary endpoints are usually correlated between them. Since the correlation affects parameters estimation and sample size calculation, we should correctly specify them, because if we define the endpoints as if they were not correlated, we will not achieve the desired power (Sozu et al., 2010).

One possible approach to deal with these challenges is to transform the multivariate problem into a univariate one combining several outcomes in a single summary indicator (Rauch et al., 2014). In this regard, it is common to use the combination of several responses into a unique variable, specially in the settings of time-to-event and binary outcomes. Then, the focus is on the time of first event between a set of possible adverse events which are assumed to be relevant to the disease progress, or, in the binary context, the composite collapses the information into a binary endpoint which takes value 1 if whenever one of the outcomes has occurred.

Composite endpoints are frequently chosen as primary endpoint in many health areas and specially in cardiovascular and oncology trials. There are three key advantages for using a composite. First, they avoid the need of multiplicity adjustment. Second, a composite endpoint contains more information on the course of the disease than a single endpoint providing a better explanation about the differences between treatment groups. Third, the increment of the number of observed events is expected to increase the power (Rauch et al., 2014).

The main drawback of using a composite endpoint is its interpretation since its components rarely have comparable clinical importance and similar treatment effects (Ferreira-González et al., 2007). Besides, a combination of different events changes the mean and also the variance of the response upon which the analysis is based (Lefkopoulou and Ryan, 1993). The addition of components which are not relevant enough could compromise the interpretation of results and reduce power. Hence, the choice of the particular components for the composite has a great importance in the design phase (Mascha and Sessler, 2011) and a deeper study about the meaning of the composite response is needed (Rauch and Kieser, 2013).

Legler et al. (1995) and Lefkopoulou and Ryan (1993) presented a framework for comparing the performance of several tests based on multiple binary endpoints. They compare as well those tests to the test that results of collapsing the data into a composite. In the framework of survival analysis, Gómez and Lagakos (2013) proposed the Asymptotic Relative Efficiency (ARE) method to compare the efficiencies of two trial designs according to the chosen primary endpoint. The motivation lied in deciding between one relevant primary endpoint or the compound of this relevant and one additional endpoint as primary endpoint of the trial. Their methodology, referred to as the ARE method, provides a tool to quantify the improvement in efficiency of adding a secondary endpoint to the primary endpoint. However, the ARE method has not been studied for binary outcomes.

Assume a binary composite endpoint and also the most severe and relevant of its components. Aiming to provide statistical guidelines that would indicate when it is more efficient to use the composite endpoint over one of its components as the primary endpoint of the trial, we expand the ARE method proposed by Gómez and Lagakos to binary endpoints. We show that the ARE method for binary composite endpoint depends on six parameters including the degree of association between the components of the composite endpoint on each group, the event proportion, and the effect of therapy on each component.

The chapter is structured as follows. In Section 3.2, we relate the parameters of the composite to the parameters of the components and the correlation between them. In Section 3.3, we find a relationship between the odds ratio of the composite and the odds ratios of its components. In Section 3.4, we define the extension of the ARE method for binary endpoints and we explain the applicability of the method. We illustrate the use of the methodology by means of a clinical trial in Section 3.5. In Section 3.6, we present recommendation guidelines in order to assess in which cases a composite endpoint should be preferred because is more efficient than the most relevant of its components. In Section 3.7 we derive the

ARE method for hypothesis problems in terms of the difference in proportions. We present the expression of the ARE for both contiguous and fixed alternatives, and discuss its relationship with the ARE definition given in terms of the odds ratios. A short discussion concludes the chapter.

3.2 Notation and main assumptions

3.2.1 Binary endpoints

Consider a randomized clinical trial comparing two treatment groups, control group ($i = 0$) and treatment group ($i = 1$), each group composed of n_i patients and denoting by $n = n_0+n_1$ the total number of patients. We assume two different binary endpoints of potential interest, ε_1 and ε_2. Let X_{ijk} denote the response of the k-th binary endpoint for the j-th patient in the i-th group of treatment ($i = 0, 1$, $j = 1, ..., n_i$, $k = 1, 2$). The response X_{ijk} is defined by 1 if the event, ε_k, has occurred and 0 otherwise. Then, for all $j \in \{1, ..., n_i\}$, $i \in \{0, 1\}$, $k \in \{1, 2\}$, the event rates are defined as:

$$p_k^{(i)} = \mathrm{P}(X_{ijk} = 1) \;=\; 1 - q_k^{(i)}$$

where $p_k^{(i)}$ and $q_k^{(i)}$ are the probabilities that ε_k occurs or not, respectively, for a patient belonging to the i-th group of treatment.

We consider a binary composite endpoint of two components, $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$, defined as the event that occurs whenever one of the endpoints is observed. Moreover, we assume that there exists one endpoint which is more relevant for the scientific question than the other, with no loss of generality, consider ε_1 the relevant endpoint and ε_2 the additional one. Denote by X_{ij*} the composite response defined as a Bernoulli random variable of parameter $p_*^{(i)} = \mathrm{P}(X_{ij*} = 1) = 1 - q_*^{(i)}$, where

$$X_{ij*} = \begin{cases} 1, \text{ if } X_{ij1} + X_{ij2} \geq 1 \\ 0, \text{ if else } X_{ij1} + X_{ij2} = 0 \end{cases}$$

In order to quantify the differences in efficacy between the two treatments, we might use the odds ratio for each k-th endpoint defined as:

$$\mathrm{OR}_k = \frac{p_k^{(1)}/q_k^{(1)}}{p_k^{(0)}/q_k^{(0)}}$$

Hereafter, we assume that both the composite endpoint and the relevant endpoint could lead to answer the clinical question, that is, both might be used as the primary endpoint of the trial.

3.2.2 The relevant endpoint as primary endpoint

If we test the treatment effect on the relevant endpoint, we establish the following hypothesis test:

$$\mathcal{H}_1 : \quad \begin{cases} H_0 : & \log(\mathrm{OR}_1) = 0 \\ H_1 : & \log(\mathrm{OR}_1) < 0 \end{cases} \tag{3.1}$$

where the null hypothesis of no-treatment effect is stated as $\mathrm{OR}_1 = 1$ or equivalently $\log(\mathrm{OR}_1) = 0$ and the alternative hypothesis assumes a risk reduction of the relevant endpoint, then, a negative $\log(\mathrm{OR}_1)$. For addressing the problem, we consider the score test defined as:

$$T_{1,n} = \frac{\hat{p}_1^{(0)} - \hat{p}_1^{(1)}}{\sqrt{\frac{1}{n_0}\hat{p}_1^{(0)}\hat{q}_1^{(0)} + \frac{1}{n_1}\hat{p}_1^{(1)}\hat{q}_1^{(1)}}} \tag{3.2}$$

where $\hat{p}_1^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij1} = 1 - \hat{q}_1^{(i)}$, that is, the proportion of relevant events in the i-th group of treatment.

Under the null hypothesis, the score test asymptotically follows the standard normal distribution. Under contiguous alternatives of the form $H_{1,n} : \log(\mathrm{OR}_1)_n = \frac{v_1}{\sqrt{n}}$, where $v_1 < 0$, the score test is asymptotically normal with unit variance and mean δ_1, called non-centrality parameter of the test, given by:

$$\delta_1 = -v_1\sqrt{p_1^{(0)}q_1^{(0)}\pi(1-\pi)} \tag{3.3}$$

where π denotes the proportion of patients allocated to control group, that is, $\pi = \lim_{n\to+\infty} n_0/n$.

For any finite sample size, $\log(\mathrm{OR}_1)_n = \frac{v_1}{\sqrt{n}}$ is the treatment effect assumed as alternative. The effects are determined by the sample size and the constant v_1 which is interpreted as the limiting treatment effect as sample size increases, i.e.:

$$\lim_{n\to+\infty} \sqrt{n}\log(\mathrm{OR}_1)_n = v_1$$

Since the contiguous alternative changes with n, it forms a sequence that converge to 0, that is, to the null hypothesis as $n \to +\infty$. Whereas the power of the score

test under any fixed alternative goes to 1 as sample size goes to infinity, under contiguous alternatives the limiting power is strictly less than 1.

3.2.3 The composite endpoint as primary endpoint

If the treatment effect is evaluated at the composite endpoint, the hypotheses are defined by:

$$\mathcal{H}_* : \quad \begin{cases} H_0 : & \log(\mathrm{OR}_*) = 0 \\ H_1 : & \log(\mathrm{OR}_*) < 0 \end{cases} \tag{3.4}$$

where under the null hypothesis, we assume that there is not treatment effect on the composite endpoint and, under the alternative hypothesis, we state a reduction of the risk evaluated on the composite event.

Following the same procedure as above, the difference between treatment groups is tested by means of the score test, $T_{*,n}$, namely:

$$T_{*,n} = \frac{\hat{p}_*^{(0)} - \hat{p}_*^{(1)}}{\sqrt{\frac{1}{n_0}\hat{p}_*^{(0)}\hat{q}_*^{(0)} + \frac{1}{n_1}\hat{p}_*^{(1)}\hat{q}_*^{(1)}}} \tag{3.5}$$

The score test $T_{*,n}$ under the null hypothesis is asymptotically $N(0,1)$, and under a sequence of contiguous alternatives, $H_{*,n} : \log(\mathrm{OR}_*)_n = \frac{v_*}{\sqrt{n}}$, is asymptotically normal with unit variance and mean δ_* (non-centrality parameter) given by:

$$\delta_* = -v_* \sqrt{p_*^{(0)} q_*^{(0)} \pi(1-\pi)} \tag{3.6}$$

3.3 Binary Composite Endpoint defined from the margins

3.3.1 Parameters

Bahadur's theorem (Bahadur, 1961) allows to determine the joint distribution of multiple correlated binary endpoints and shows that the joint distribution is uniquely determined by the marginal probabilities and the degree of association between the endpoints. As noted by Sozu et al. (2010), the association degree among the correlated binary endpoints might be defined by different measures. We consider *Pearson's correlation coefficient* as the association measure between

endpoints. Let $\rho^{(i)}$ be the correlation coefficient, also referred as phi coefficient, in the i-th group defined as:

$$\rho^{(i)} = \frac{p_{\cap}^{(i)} - p_1^{(i)} p_2^{(i)}}{\sqrt{p_1^{(i)} q_1^{(i)} p_2^{(i)} q_2^{(i)}}}$$

Note that the correlation coefficient is represented by the underlying probabilities and the overlap between these marginal events, expressed by $p_{\cap}^{(i)} = \mathrm{P}(X_{ij1} = 1, X_{ij2} = 1)$. Applying results from Bahadur (1961), the probability of the composite endpoint in the i-th group of treatment, $p_*^{(i)}$, is uniquely determined by the probabilities of the single endpoints, $p_1^{(i)}$, $p_2^{(i)}$, and the correlation between them, as follows:

$$p_*^{(i)} = 1 - q_1^{(i)} q_2^{(i)} - \rho^{(i)} \sqrt{p_1^{(i)} p_2^{(i)} q_1^{(i)} q_2^{(i)}}\ , \qquad i = 0, 1 \tag{3.7}$$

The odds ratio of the composite endpoint, OR_*, is given in terms of the correlation between endpoints for each group, the event proportions in the control group given by the respective odds and the therapy effect given by the corresponding odds ratio, as follows:

$$\mathrm{OR}_* = \frac{\dfrac{\left(1+\frac{\mathrm{OR}_1 p_1^{(0)}}{1-p_1^{(0)}}\right)\left(1+\frac{\mathrm{OR}_2 p_2^{(0)}}{1-p_2^{(0)}}\right) - 1 - \rho^{(1)}\sqrt{\frac{\mathrm{OR}_1 \mathrm{OR}_2 p_1^{(0)} p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho^{(1)}\sqrt{\frac{\mathrm{OR}_1 \mathrm{OR}_2 p_1^{(0)} p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}}{\dfrac{\left(1+\frac{p_1^{(0)}}{(1-p_1^{(0)})}\right)\left(1+\frac{p_2^{(0)}}{(1-p_2^{(0)})}\right) - 1 - \rho^{(0)}\sqrt{\frac{p_1^{(0)} p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho^{(0)}\sqrt{\frac{p_1^{(0)} p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}} \tag{3.8}$$

The full derivation is to be found in Appendix A.1. Observe that OR_* depends on the following six parameters $(p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho^{(0)}, \rho^{(1)})$ and that the parameters associated to each component together with the correlation between them is what we only need to assess the effect on the composite endpoint.

A special property of binary endpoints is that the correlation takes values between two bounds which are defined according to the marginal probabilities (Prentice, 1988) –the parametric space of $\rho^{(i)}$ is more confined than $(-1, 1)$–, that is:

$$\rho^{(i)} \in [\ m(p_1^{(i)}, p_2^{(i)}),\quad M(p_1^{(i)}, p_2^{(i)})\ ] \subseteq [-1, 1]$$

where:

$$m(p_1^{(i)}, p_2^{(i)}) = \max\left\{ -\sqrt{\frac{p_1^{(i)} \cdot p_2^{(i)}}{q_1^{(i)} \cdot q_2^{(i)}}}, -\sqrt{\frac{q_1^{(i)} \cdot q_2^{(i)}}{p_1^{(i)} \cdot p_2^{(i)}}} \right\}$$

$$M(p_1^{(i)}, p_2^{(i)}) = \min\left\{ +\sqrt{\frac{p_1^{(i)} \cdot q_2^{(i)}}{p_2^{(i)} \cdot q_1^{(i)}}}, +\sqrt{\frac{p_2^{(i)} \cdot q_1^{(i)}}{p_1^{(i)} \cdot q_2^{(i)}}} \right\}$$

3.3.2 Treatment effects and non-equivalence between hypotheses

It can be easily proved by inspection of (3.8) that if the treatment has no effect on any of the marginal components and the correlation between them is the same in the two groups, then, the treatment has no effect on the composite endpoint, that is:

$$\mathrm{OR}_1 = 1, \quad \mathrm{OR}_2 = 1, \quad \rho^{(0)} = \rho^{(1)} \Longrightarrow \quad \mathrm{OR}_* = 1$$

Note that this result could be restated in terms of the event proportions:

$$p_1^{(0)} = p_1^{(1)}, \quad p_2^{(0)} = p_2^{(1)}, \quad \rho^{(0)} = \rho^{(1)} \quad \Longrightarrow \quad p_*^{(0)} = p_*^{(1)}$$

However, the reciprocal is not necessarily true, that is to say, the effect of treatment on any endpoint ($\mathrm{OR}_1 < 1$ or $\mathrm{OR}_2 < 1$) could be diluted on the composite ($\mathrm{OR}_* = 1$). This complex relationship between the odds ratios of each component and the composite shows how the treatment effects are differently measured on each endpoint, and cannot be taken as equivalent. Thus, the two hypothesis tests being considered to test the treatment effect on either endpoint, $\mathcal{H}_1$ (stated in (3.1)) and $\mathcal{H}_*$ (stated in 3.4)), are not equivalent.

3.4 Asymptotic Relative Efficiency

We extend the ARE method developed by Gómez and Lagakos for time-to-event endpoints to binary endpoints. The extension relies on the asymptotic behaviours of the score tests $T_{1,n}$ for the relevant endpoint given in (3.2) and $T_{*,n}$ for the composite endpoint given in (3.5) presented in section 3.2, instead of the log-rank test that was used for survival endpoints. In the following sections we present the ARE method and its version for fixed alternatives.

3.4.1 ARE method for contiguous alternatives

Consider the following not equivalent hypothesis tests based on the relevant endpoint and on the composite endpoint:

$$\mathcal{H}_{1,n}: \quad \begin{cases} H_0: & \log(\mathrm{OR}_1) = 0 \\ H_{1,n}: & \log(\mathrm{OR}_1)_n = \frac{v_1}{\sqrt{n}} \end{cases} \qquad \mathcal{H}_{*,n}: \quad \begin{cases} H_0: & \log(\mathrm{OR}_*) = 0 \\ H_{*,n}: & \log(\mathrm{OR}_*)_n = \frac{v_*}{\sqrt{n}} \end{cases} \tag{3.9}$$

Let $T_{1,n}$, $T_{*,n}$ be the score tests corresponding to $\mathcal{H}_{1,n}$ and $\mathcal{H}_{*,n}$, respectively. Whereas under the null hypothesis both tests asymptotically follow the standard normal distribution, under contiguous alternatives, they are asymptotically $N(\delta_1, 1)$ and $N(\delta_*, 1)$ with δ_1 and δ_* presented in (3.3) and (3.6), respectively. Both tests behave as a displaced normal distribution according to the non-centrality parameter of the test, δ_1 and δ_*. Since the power of both tests is governed by the non-centrality parameters δ_1 and δ_*, and the larger the parameter is the greater the power (see Figure 3.1), a comparison between them yields a criterion for relative efficiency. We define the ARE as the square of the ratio of the non-centrality parameters, that is:

$$\mathrm{ARE}(T_{*,n}, T_{1,n}) \;=\; \left(\frac{\delta_*}{\delta_1}\right)^2 \;=\; \frac{v_*^2 p_*^{(0)} q_*^{(0)}}{v_1^2 p_1^{(0)} q_1^{(0)}}. \tag{3.10}$$

$\mathrm{ARE}(T_{*,n}, T_{1,n}) > 1$ would imply larger powers if using ε_* while $\mathrm{ARE}(T_{*,n}, T_{1,n}) \leq 1$ would be in favour of using ε_1 as the best option for primary endpoint. Hence, choosing between ε_1 or ε_* is reduced to a comparison between the two means of the asymptotic law under contiguous alternatives. The best primary endpoint would be the one which has the greatest non-centrality parameter.

The method quantifies the differences in efficiency of using the composite or the relevant as primary endpoint to lead the trial and, moreover, provides a decision rule to define the primary endpoint. If the ARE is larger than 1, the composite endpoint may be considered the best option as primary endpoint. Otherwise, the relevant endpoint is preferred. However, when the ARE value is in the vicinity of one, the advantages of the composite endpoint over the relevant endpoint are too small to counteract the complicate interpretation of the composite endpoint. Thus, under this circumstance, the relevant endpoint could be used instead as primary endpoint.

Summarizing, for every endpoint, given their event rates in the control group and their limiting treatment effect, the ARE value captures which endpoint is

more efficient for designing a clinical trial and provides a criterion to choose among them.

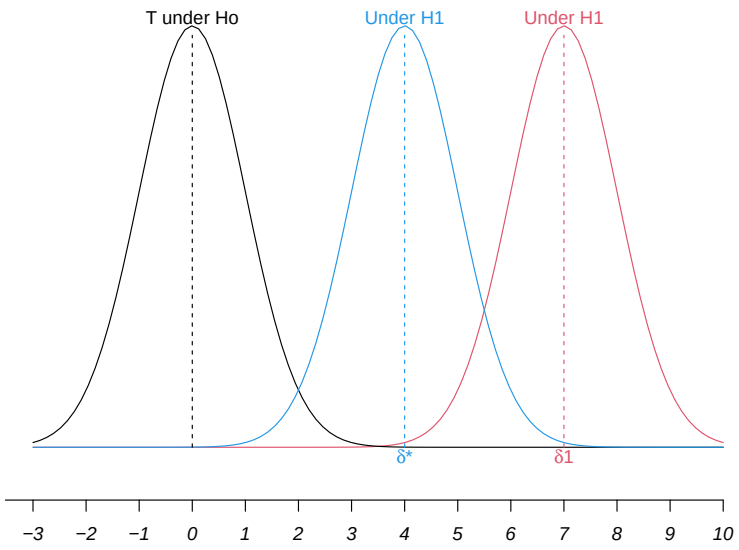


Fig. 3.1 Asymptotic behavior of the score test under the null hypothesis (most left curve) and under contiguous alternatives for each endpoint ε_1 (most right curve) and ε_* (second right).

3.4.2 ARE method for fixed alternatives

The ARE criterion to choose the primary endpoint given in (3.10) is based on alternative odds ratios which are close to 1. From a practical point of view, the interest might often be on detecting treatment effects OR_k $(k = 1, *)$, not necessarily near 1, and to address this we propose an approximated ARE value.

The efficiency criterion for fixed alternatives, based on the two-sample score tests $T_{*,n}$ and $T_{1,n}$, is defined as:

$$are(p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2, \rho^{(0)}, \rho^{(1)}) = \frac{(\log(\mathrm{OR}_*))^2 p_*^{(0)} q_*^{(0)}}{(\log(\mathrm{OR}_1))^2 p_1^{(0)} q_1^{(0)}} \tag{3.11}$$

Expression (3.11) approaches the ARE definition in (3.10) if for each endpoint, we would consider the fixed treatment effect stabilized for the sample size as an approximate value for the limiting treatment, that is: $\sqrt{n}\log(\mathrm{OR}_1) \cong v_1$ and $\sqrt{n}\log(\mathrm{OR}_*) \cong v_*$.

Hence, the decision on whether to use a composite binary endpoint versus its most relevant component as the primary endpoint can be assessed by computing the ARE for fixed alternatives, referred to as *are*. The *are* depends on the joint law of (X_{ij1}, X_{ij2}) $(i = 0, 1)$ and can be determined by the following anticipated parameters: (i) $p_1^{(0)}$ and $p_2^{(0)}$, event rates in control group for the relevant endpoint, ε_1, and the additional endpoint, ε_2; (ii) OR_1 and OR_2 fixed treatment effects for ε_1 and ε_2; (iii) $\rho^{(0)}$ and $\rho^{(1)}$, correlation between X_{ij1} and X_{ij2} for each group.

3.5 TAXUS-V trial

Drug-eluting stents have been proved to reduce restenosis in noncomplex lesions, even so, their utility has not been studied in a patient population with more complex lesions. TAXUS-V was a prospective, multicenter, randomized trial to investigate the safety and efficacy of a paclitaxel-eluting stent in a patient population with more complete lesions than previously studied (Stone et al., 2005). The trial was conducted from February 2003 to March 2004 at 66 academic and community-based institutions with 1156 patients who underwent stent implantation in a single coronary artery stenosis, including 664 patients (57.4%) with complex or previously unstudied lesions and 9-month clinical and angiographic follow-up. Patients were randomly assigned to receive one or more bare metal stents ($n = 579$) or identical-appearing paclitaxel-eluting stents ($n = 577$).

The primary endpoint was the 9-month incidence of ischemia-driven target vessel revascularization, ε_1. As a secondary endpoint, major adverse cardiac events, ε_*, were defined as ischemia-driven target-vessel revascularization, ε_1, or death from cardiac causes or myocardial infarction, ε_2. The study shows that compared with a bare metal stent, implantation of the paclitaxeleluting stent in a patient population with complex lesions effectively reduces the rate of vessel revascularization.

For illustrative purposes, we assume that a study in a similar setting is to be planned, and the question that arises is which primary endpoint should be used to lead the trial. We also assume that the results of TAXUS-V are used for this purpose. Aiming to study whether it would be more efficient to base the

study on major adverse cardiac events, ε_*, instead of ischemia-driven target vessel revascularization, ε_1, we exemplify the use of the ARE method.

The frequency of target vessel revascularization in bare metal group is $p_1^{(0)} = 0.173$, whereas the frequency of death from cardiac causes or myocardial infarction is $p_2^{(0)} = 0.055$. Furthermore, the frequencies under the test group are $p_1^{(1)} = 0.121$ and $p_2^{(1)} = 0.057$, respectively. We discuss the use of the composite endpoint as primary endpoint, for given values $p_1^{(0)} = 0.173$, $p_2^{(0)} = 0.055$ and $p_1^{(1)} = 0.121$ and for the values for the parameter $p_2^{(1)}$ and ρ presented in Table 3.1. For given pairs $(p_1^{(0)}, p_2^{(0)})$ and $(p_1^{(1)}, p_2^{(1)})$ and assuming equal correlation in both groups, the eligible values for ρ lie in the interval $(-0.09, 0.53)$.

Table 3.1 Values of $p_1^{(0)}$ and $p_1^{(1)}$, probability of target vessel revascularization in bare metal group and in paclitaxel-eluting group; $p_2^{(0)}$ and $p_2^{(1)}$, probability of death from cardiac causes or myocardial infarction in bare metal group and in paclitaxel-eluting group; ρ, correlation among target vessel revascularization and death from cardiac causes or myocardial infarction; OR_1, odds ratio for target vessel revascularization; OR_2, odds ratio for death from cardiac causes or myocardial infarction, used for the discussion. The left part of the table shows the treatment effects in terms of p, the right part shows the treatment effects in terms of OR.

Parameter	Values	Parameter	Values
$p_1^{(0)}$	0.173	$p_1^{(0)}$	0.173
$p_1^{(1)}$	0.121	OR_1	0.67
$p_2^{(0)}$	0.055	$p_2^{(0)}$	0.055
$p_2^{(1)}$	$0.057, 0.050, 0.045, 0.040, 0.035$	OR_2	$1.04, 0.90, 0.81, 0.72, 0.62$
ρ	$(-0.09, 0.53)$	ρ	$(-0.09, 0.53)$

Figure 3.2 depicts the ARE values (in log scale) in terms of the correlation for each of the five different values of the treatment effect on ε_2. Observe that for a fixed correlation the ARE takes greater values as the odds ratio OR_2 for death from cardiac causes or myocardial infarction decreases. Therefore, the composite endpoint becomes more effective and more useful when the odds ratio for the additional endpoint, OR_2, shows a greater treatment effect. Furthermore, notice that the ARE decreases when the correlation increases, that is, the more correlated among target vessel revascularization and death from cardiac causes or myocardial infarction are, the less appropriate necessary is the composite as primary endpoint.

Especially, when the odds ratio for death from cardiac causes or myocardial infarction, OR_2, is equal or larger than 0.81, the ARE is almost always less than 1 (see Figure 3.2). Hence, the use of target vessel revascularization, ε_1, provide more efficient detection of the differences between treatments. In the case that $OR_2 \leq 0.62$, the ARE is greater than 1. Then, the primary endpoint major adverse cardiac events, ε_*, would have been more efficient instead of relevant endpoint. Finally, note that when OR_2 is around 0.72, the decision depends on the value that correlation has.

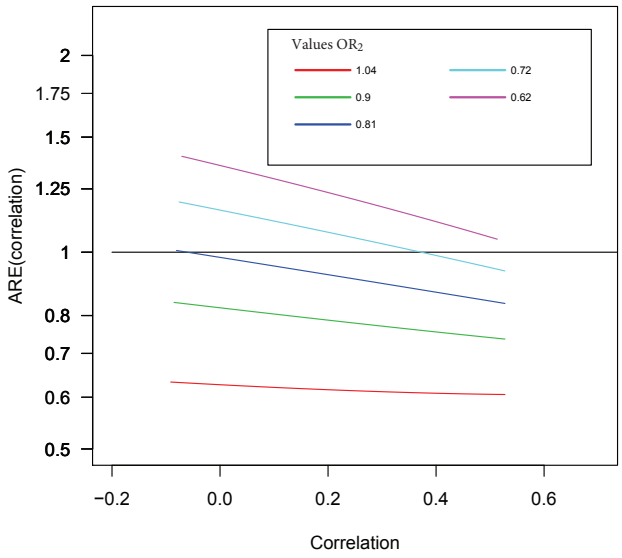


Fig. 3.2 ARE of major adverse cardiac events (death from cardiac causes, myocardial infarction, or target-vessel revascularization) versus target-vessel revascularization for a range of correlation coefficient and different values of OR_2 for the parameters: $p_1^{(0)} = 0.173$, $p_2^{(0)} = 0.055$ and $p_1^{(1)} = 0.121$. The plot shows the curves of the *are* for each OR_2 depending on the assumed ρ.

3.6 Statistical efficiency guidelines

We have seen that the relative efficiency to choose between a composite endpoint or one of its relevant components can be expressed in terms of the following anticipated parameters: treatment effects, event rates and correlations. In this section, we discuss the influence that these parameters have on the relative efficiency value. For example, which role does play the correlation between the two components in preferring the composite as primary endpoint? We conclude reporting guidelines which could be of some help when designing a randomized clinical trial and facing the choice between several binary endpoints or their combination.

3.6.1 Design

Our efficiency guidelines will be based on event rates, $p_1^{(0)}$, $p_2^{(0)}$, smaller than 0.1, odds ratios, OR_1, OR_2, between 0.5 and 1, and positive correlations (see Table 3.2). This choice is in accordance with the values that are usually encountered in clinical trials. From now on, we assume that the correlations are the same in the two groups and we denote it by ρ. Although Table 3.2 yields 436810 scenarios, since for every pair $(p_1^{(0)}, p_2^{(0)})$ and $(p_1^{(1)}, p_2^{(1)})$ not all the correlation values are feasible, the total number of possible scenarios is reduced to 315348.

Since the ARE method for fixed alternatives given in (3.11) depends on the parameters $p_1^{(0)}$, $p_2^{(0)}$, OR_1, OR_2 and ρ, we calculate the *are* for each scenario. The ARE values that we have obtained has 1.52 as a median, and 0.81 and 4.82 as first and third quartile. We follow the principle that if $are > 1$, the use of the composite endpoint is recommended, and if $are \leq 1$, the relevant endpoint should be used as primary endpoint. At last, we compute the percentage of cases on which the composite is preferred over the relevant endpoint. We conclude with recommendations for the choice of the primary endpoint in terms of the values of the correlations, the treatment effects and the event rates in control group for each individual component. We have performed all computations using `R` software tool (Version 0.98.1087), the time required to perform the considered scenarios was 16.58h.

As said earlier, when the ARE values are close to one, in particular if $are \in (1, 1.1)$, the benefits of using the composite endpoint over the relevant endpoint are small. Despite the value one is regarded as the threshold of our study and is

the focus of the subsequent discussion, guidelines using 1.1 as the threshold for the decision can be viewed in Appendix B.1.

Table 3.2 Values of parameters $p_1^{(0)}, p_2^{(0)}, \mathrm{OR}_1, \mathrm{OR}_2$ and ρ for the settings used for the efficient guidelines.

Parameter	Values
$p_1^{(0)}$, $p_2^{(0)}$	$0.010, 0.015, 0.020, 0.025, 0.030, 0.035, 0.040, 0.045, 0.050,$ $0.055, 0.060, 0.065, 0.070, 0.075, 0.080, 0.085, 0.090, 0.095, 0.100$
OR_1, OR_2	$0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.99$
ρ	$0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$
Total scenarios	436810
Possible scenarios	315348

3.6.2 General pattern of the percentage of cases in which $are > 1$

We study the influence that the value of certain anticipated parameters, such as the treatment effect on the relevant endpoint or the event rate of the additional endpoint, has on the selection between a composite endpoint or its more relevant component as primary endpoint. As we will see, the most well-suited primary endpoint might differ according to the anticipated parameters of the clinical trial.

We have computed the *are* values for each of the 315348 scenarios described in Table 3.2 and in each case we have recorded whether $are > 1$ –the composite endpoint would be recommended– or $are \leq 1$ –the relevant endpoint should be kept as primary endpoint–. A given scenario is characterized by the following 5 parameter values $\theta = (\mathrm{OR}_1, \mathrm{OR}_2, \rho, p_1^{(0)}, p_2^{(0)})$. Let $P_1(a)$ indicate the percentage of cases yielding $are > 1$ among all the scenarios with $\mathrm{OR}_1 = a$. Analogously define $P_j(a)$ as the percentage of cases yielding $are > 1$ among all the scenarios with $\theta_j = a$ $(j = 2, \cdots, 5)$.

We have examined $P_1(\mathrm{OR}_1)$ for $0.5 \leq \mathrm{OR}_1 < 1$, $P_2(\mathrm{OR}_2)$ for $0.5 \leq \mathrm{OR}_2 < 1$ and $P_3(\rho)$ for $0 \leq \rho \leq 1$. We observe that the percentage of situations in which $are > 1$ increases whenever: i) the relative effect of treatment on the relevant endpoint increases, ii) the relative effect of treatment on the additional endpoint decreases and iii) the correlation between the two endpoints decreases. In other words, the number of situations where the composite endpoint is preferred is larger i) for

larger values of OR_1, ii) for smaller values of OR_2 and iii) for weakly correlated endpoints. Figure 3.3 and Figures B.1, B.2 (in Appendix B.1) summarize these findings.

We have studied the behavior of $P_2(\mathrm{OR}_2)$ as a function of OR_1. Figure 3.4 represents $P_2(\mathrm{OR}_2 = \mathrm{OR}_1 + a)$ for $\mathrm{OR}_1 = 0.6$ and $-0.10 \leq a \leq 0.35$. We observe that the percentage of cases in which the composite is preferred drops off rapidly when the effect of treatment is not as strong on the additional endpoint as it is on the relevant endpoint (see Figures B.3 and B.4 when $\mathrm{OR}_1 = 0.7$ and $\mathrm{OR}_1 = 0.8$ in Appendix B.1).

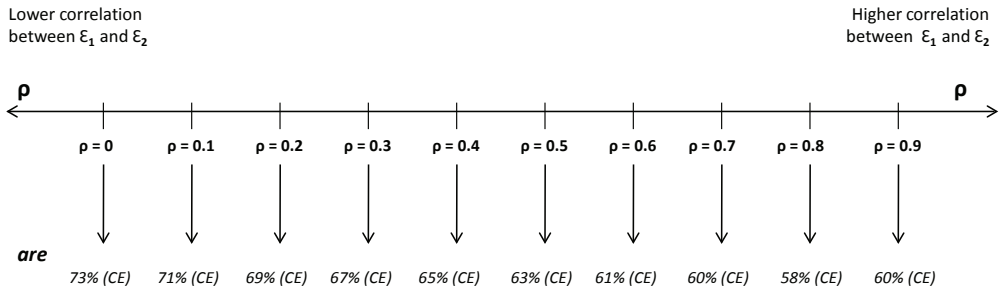


Fig. 3.3 Percentage of scenarios in which the composite endpoint should be used depending on ρ.

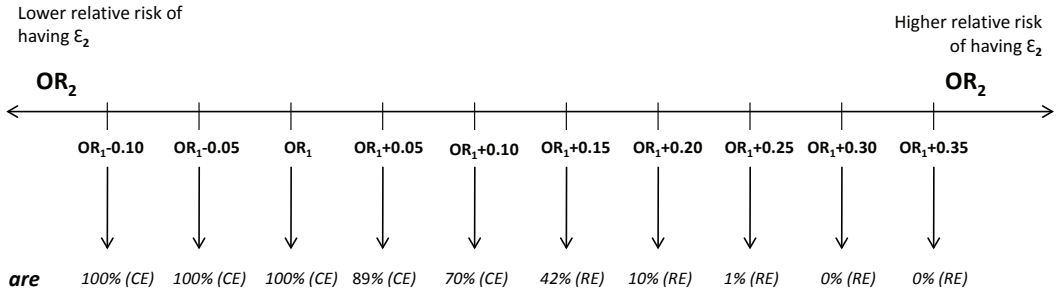


Fig. 3.4 Percentage of scenarios in which the composite endpoint should be used depending on OR_2 when $\mathrm{OR}_1 = 0.6$.

We have also studied the behavior of $P_4(p_1^{(0)})$ for $0.01 \leq p_1^{(0)} \leq 0.1$ and of $P_5(p_2^{(0)})$ for $0.01 \leq p_2^{(0)} \leq 0.1$. There is a certain trend showing (plots not provided) that

$P_4(p_1^{(0)})$ decreases with $p_1^{(0)}$ while $P_5(p_2^{(0)})$ increases with $p_2^{(0)}$, indicating that less frequent event rates for the relevant endpoint and more frequent for the additional endpoint are in the direction of preferring the composite endpoint. However, the values for $P_4(p_1^{(0)})$ or $P_5(p_2^{(0)})$ are between 60% and 75%, implying that the other parameters $(\mathrm{OR}_1, \mathrm{OR}_2, \rho)$ play a more important role in the choice between the relevant and the composite endpoint.

3.6.3 Recommendations for the choice of the primary endpoint

We have split the recommendations into the following three cases: (I) when the correlation takes values between 0 and 1 ($0 < \rho < 1$); (II) when the relevant and the additional endpoint are independent ($\rho = 0$); and (III) when $\rho = 1$, implying that the relevant and the additional endpoint take the same value.

(I) Although the total number of scenarios that we have reproduced is very large and it has been useful to understand how the *are* behaves, when it comes to anticipate parameter values on which to base our decisions, accuracy cannot be as slim and is more realistic to render the recommendations to 3 or 4 categories of association, of strengths of the relative effect and of levels of frequency of the events. To this end, we have chosen four degrees of association: weak ($0 < \rho < 0.3$), medium-weak ($0.3 \leq \rho < 0.6$), medium-strong ($0.6 \leq \rho < 0.8$), strong ($0.8 \leq \rho < 1$); three categories for treatment effect: Large for Odds Ratios between 0.5 and 0.7, Medium for Odds Ratios between 0.7 and 0.9 and Low for Odds Ratios between 0.9 and 1; and four event rates in control group for the relevant and additional endpoints, low ($p \leq 0.025$), medium-low ($0.025 \leq p \leq 0.05$), medium-large ($0.05 \leq p \leq 0.075$), large ($p > 0.075$).

To derive recommendations, for each case we provide the percentage of cases in which the composite is preferred. On the basis of these percentages, we indicate whether the relevant or composite endpoint should be used. We are considering here that if the percentage of $are > 1$ is larger than 60%, the recommendation is to use of the composite endpoint; if the percentage is less than 40%, the recommendation is to use of the relevant endpoint; otherwise, if the percentage lies between 40% and 60%, the recommendation cannot be given. In this last case, we have reported that the recommendation is not conclusive and we have written CE/RE. There are not conclusions for all situations, therefore, the ensuing computation of the ARE is needed for the rest of particular situations.

Table 3.3 summarizes the recommendation in terms of the categories for $(\mathrm{OR}_1, \mathrm{OR}_2)$. Basically, the composite endpoint should be used when: i) treatment

effect on the additional endpoint is large; ii) treatment effects on the relevant and additional endpoint are medium; iii) treatment effects on the relevant and additional endpoint are low and medium, respectively. On the other hand, the relevant endpoint is almost always preferred if the treatment effect on the additional endpoint is low and the treatment effect on the relevant is large or medium.

Table 3.3 Recommendations in terms of treatment effects of the relevant and the additional endpoint, large ($0.5 \leq \text{OR} < 0.7$), medium ($0.7 \leq \text{OR} < 0.9$) or low ($0.9 \leq \text{OR} < 1$). Each cell indicates whether the relevant endpoint (RE) ($are \leq 1$) or composite endpoint (CE) ($are > 1$) should be used and, in parentheses, the percentage of cases in which composite is preferred based on the scenarios described in Table 3.2.

	Large treatment effect on ε_2	Medium treatment effect on ε_2	Low treatment effect ε_2
Large treatment effect on ε_1	CE (91.18%)	RE (23.06%)	RE (0%)
Medium treatment effect on ε_1	CE (100%)	CE (83.65%)	RE (6.52%)
Low treatment effect ε_1	CE (100%)	CE (100%)	CE (68.81%)

Recommendations taking into account the level of association together with the treatment effects on the relevant and on the additional endpoint, event rates in control group for both the relevant and the additional endpoint are summarized in Table 3.4. As earlier, we observe that the percentage of $are > 1$ decreases as the degree of association increases. This underlines the importance of the correlation to decide the primary endpoint. In particular, when the treatment effect either on the relevant or additional endpoint is medium, the value of the correlation might play a crucial role in the decision. Notice that the percentages of $are > 1$ in terms of the event rates are never larger than 75% or smaller than 50%, hence the frequency of the relevant and additional endpoints cannot characterize by themselves the decision on which primary endpoint to use.

(II) Whenever the relevant and additional endpoints are independent ($\rho = 0$), the composite endpoint would be intuitively preferred, however this is not always the case as Figure 3.3 shows. Following the rationale outlined above, Table 3.5 takes care of this situation. Note that the relevant endpoint is always preferred to the composite endpoint when the treatment effect on the relevant endpoint is large and the treatment effect on the additional endpoint is low. Besides, whenever the treatment effect on the relevant endpoint is medium and the treatment effect on the additional endpoint is low, the relevant endpoint should be the primary

Table 3.4 Recommendations in terms of degree of association between endpoints, weak ($0 < \rho < 0.3$), medium-weak ($0.3 \leq \rho < 0.6$), medium-strong ($0.6 \leq \rho < 0.8$), strong ($0.8 \leq \rho < 1$); treatment effects of the relevant and the additional endpoint, large ($0.5 \leq \text{OR} < 0.7$), medium ($0.7 \leq \text{OR} < 0.9$) or low ($0.9 \leq \text{OR} < 1$); event rates in control group for the relevant and additional endpoints, low ($p \leq 0.025$), medium-low ($0.025 \leq p \leq 0.05$), medium-large ($0.05 \leq p \leq 0.075$), large ($p > 0.075$). Each cell indicates whether the relevant endpoint (RE) ($are \leq 1$) or composite endpoint (CE) ($are > 1$) should be used and, in parentheses, the percentage of cases in which composite is preferred based on the scenarios described in Table 3.2.

	Correlation			
	Weak	Medium-weak	Medium-strong	Strong
Large treatment effect on ε_2	CE (99.72%)	CE (97.41%)	CE (92.87%)	CE (84.97%)
Medium treatment effect on ε_2	CE (74.96%)	CE (65.97%)	CE/RE (58.23%)	CE/RE (56.96%)
Low treatment effect ε_2	RE (23.61%)	RE (21.39%)	RE (20.99%)	RE (28.16%)
Large treatment effect on ε_1	CE/RE (49.80%)	CE/RE (42.29%)	RE (35.72%)	RE (38.00%)
Medium treatment effect on ε_1	CE (73.47%)	CE (68.72%)	CE (63.04%)	CE/RE (57.78%)
Low treatment effect ε_1	CE (92.16%)	CE (91.05%)	CE (89.87%)	CE (86.61%)
Low event rate for ε_1	CE (74.80%)	CE (73.89%)	CE (68.66%)	CE (65.36%)
Medium-low event rate for ε_1	CE (70.68%)	CE (66.41%)	CE (66.05%)	CE (62.95%)
Medium-large event rate for ε_1	CE (68.32%)	CE (62.78%)	CE/RE (59.01%)	CE (62.14%)
Large event rate for ε_1	CE (67.00%)	CE (60.52%)	CE/RE (53.29%)	CE/RE (50.22%)
Low event rate for ε_2	CE (66.63%)	CE/RE (57.96%)	CE/RE (53.87%)	CE/RE (52.92%)
Medium-low event rate for ε_2	CE (69.38%)	CE (64.16%)	CE/RE (55.77%)	CE/RE (54.99%)
Medium-large event rate for ε_2	CE (71.00%)	CE (67.34%)	CE (61.66%)	CE/RE (55.46%)
Large event rate for ε_2	CE (72.16%)	CE (69.20%)	CE (66.87%)	CE (66.61%)

endpoint to lead the trial. Otherwise, if the treatment effect on the additional endpoint is large ($OR_2 \leq 0.7$), the composite endpoint is always preferred.

Table 3.5 Recommendations in case of independence between the relevant and the additional endpoint ($\rho = 0$) in terms of treatment effects of the relevant and the additional endpoint, large ($0.5 \leq OR < 0.7$), medium ($0.7 \leq OR < 0.9$) or low ($0.9 \leq OR < 1$). Each cell indicates whether the relevant endpoint (RE) ($are \leq 1$) or composite endpoint (CE) ($are > 1$) should be used and, in parentheses, the percentage of cases in which composite is preferred based on the scenarios described in Table 3.2.

	Large treatment effect on ε_2	Medium treatment effect on ε_2	Low treatment effect ε_2
Large treatment effect on ε_1	CE (100%)	CE/RE (48.84%)	RE (0%)
Medium treatment effect on ε_1	CE (100%)	CE (96.36%)	RE (15.12%)
Low treatment effect ε_1	CE (100%)	CE (100%)	CE (76.55%)

(III) The case of $\rho = 1$ was excluded from the settings of scenarios because in this case $are = 1$. The reason is that perfect linear dependence implies that the probabilities of the composite and the relevant endpoint are the same. As a result, it can be seen by inspection of (3.11) that the resulting are is equal to one. Hence, the decision rule sets up an equivalence between the relevant and composite endpoints in terms of efficiency.

3.7 Further work

In this section we derive the ARE for hypothesis problems in terms of the difference in proportions. We present the ARE for both contiguous and fixed alternatives, and discuss its relationship with the ARE in terms of the odds ratios given in (3.10).

3.7.1 An extension of the ARE method for difference in proportions

As before, we assume two different binary endpoints of potential interest, ε_1 and ε_2, and the binary composite endpoint $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$. Additionally, we assume that

there exists one endpoint which is more relevant for the scientific question than the other. Consider ε_1 the relevant endpoint and ε_2 the additional one.

Hypothesis problem in terms of the difference in proportions

If we consider the relevant endpoint ε_1 as primary endpoint, we can formulate the hypothesis problem as:

$$\mathcal{H}_1^{(p)} := \quad \begin{cases} H_0: & p_1^{(0)} - p_1^{(1)} = 0 \\ H_1: & p_1^{(0)} - p_1^{(1)} > 0 \end{cases}$$

where, as before, the null hypothesis states the non-treatment effect $p_1^{(0)} = p_1^{(1)}$ and the alternative hypothesis assumes a risk reduction in the intervention group as compared with the control group.

Let $T_{1,n}$ be the statistic defined in (3.2), that is:

$$T_{1,n} = \frac{\hat{p}_1^{(0)} - \hat{p}_1^{(1)}}{\sqrt{\frac{1}{n_0}\hat{p}_1^{(0)}\hat{q}_1^{(0)} + \frac{1}{n_1}\hat{p}_1^{(1)}\hat{q}_1^{(1)}}}$$

where $\hat{p}_1^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij1} = 1 - \hat{q}_1^{(i)}$. The statistic $T_{1,n}$ is asymptotically $N(0,1)$ under H_0. For any fixed alternative, $T_{1,n}$ is consistent so that the power of $T_{1,n}$ will go to 1 as $n \to +\infty$. Let us consider the sequence of contiguous alternatives to H_0 defined by:

$$H_{1,n}: \; p_1^{(0)} - p_{1,n}^{(1)} = \frac{u_1}{\sqrt{n}}$$

where $u_1 \in \mathbb{R}^+$. Observe that $p_1^{(0)}$ is considered fixed and $p_{1,n}^{(1)}$ varies depending on the sample size, n. For any finite n, $p_1^{(0)} - p_{1,n}^{(1)} = \frac{u_1}{\sqrt{n}}$ is the treatment effect assumed as alternative. Besides, the constant u_1 is interpreted as the limiting treatment effect as $n \to +\infty$, i.e.:

$$\sqrt{n}(p_1^{(0)} - p_{1,n}^{(1)}) \longrightarrow u_1$$

Under the sequences of contiguous alternatives $H_{1,n}$, $T_{1,n}$ is asymptotically normal with unit variance and mean (non-centrality parameter) μ_1 defined as:

$$\mu_1 = u_1\sqrt{\frac{\pi(1-\pi)}{p_1^{(0)}q_1^{(0)}}}$$

If the treatment is evaluated through the composite endpoint ε_*, the hypothesis problem is stated as:

$$\mathcal{H}_*^{(p)} := \quad \begin{cases} H_0: & p_*^{(0)} - p_*^{(1)} = 0 \\ H_1: & p_*^{(0)} - p_*^{(1)} > 0 \end{cases}$$

where under the null hypothesis, we assume that there is not treatment effect on the composite endpoint and, under the alternative hypothesis, there is a reduction in the number of composite events in the intervention group as compared with the control group.

Let $T_{*,n}$ be the statistic defined in (3.5), that is:

$$T_{*,n} = \frac{\hat{p}_*^{(0)} - \hat{p}_*^{(1)}}{\sqrt{\frac{1}{n_0}\hat{p}_*^{(0)}\hat{q}_*^{(0)} + \frac{1}{n_1}\hat{p}_*^{(1)}\hat{q}_*^{(1)}}}$$

The score test $T_{*,n}$ is asymptotically $N(0,1)$ under $H_0 : p_*^{(0)} - p_*^{(1)} = 0$. Under sequences of contiguous alternatives of the form $H_{*,n} : \; p_*^{(0)} - p_{*,n}^{(1)} = \frac{u_*}{\sqrt{n}}$, the statistic $T_{*,n}$ is asymptotically $N(\mu_*, 1)$, where the non-centrality parameter μ_* is given by:

$$\mu_* = u_* \sqrt{\frac{\pi(1-\pi)}{p_*^{(0)} q_*^{(0)}}} \tag{3.12}$$

ARE method for difference in proportions

Consider the following hypothesis tests based on the relevant endpoint, ε_1, and on the composite endpoint, ε_*, in terms of the difference in proportions:

$$\mathcal{H}_{1,n}^{(p)} := \quad \begin{cases} H_{0,1}: & p_1^{(0)} = p_1^{(1)} \\ H_{1,1}: & p_{1,n}^{(1)} = p_1^{(0)} - \frac{u_1}{\sqrt{n}} \end{cases} \qquad \mathcal{H}_{*,n}^{(p)} := \quad \begin{cases} H_{0,*}: & p_*^{(0)} = p_*^{(1)} \\ H_{1,*}: & p_{*,n}^{(1)} = p_*^{(0)} - \frac{u_*}{\sqrt{n}} \end{cases} \tag{3.13}$$

Let $T_{1,n}, T_{*,n}$ be the statistics given in (3.2) and (3.5) for testing $\mathcal{H}_{1,n}^{(p)}$ and $\mathcal{H}_{*,n}^{(p)}$, respectively. Note that both tests asymptotically follow the standard normal distribution under the null hypothesis of non-treatment effect, while under the alternative hypothesis asymptotically follow a normal distribution with variance 1 and non-centrality parameter μ_1 and μ_*, respectively.

To assess the difference in efficiency between ε_1 and ε_*, we base our decision on the comparison between the asymptotic behavior of $T_{1,n}$ under $\mathcal{H}_1$ and $T_{*,n}$ under $\mathcal{H}_*$. Following the same rationale that in section 3.4, we define the ARE as the square of the ratio of the non-centrality parameters as a criterion for relative efficiency. The ARE is then defined by the square of the ratio of μ_*, μ_1, that is:

$$\mathrm{ARE}_p(T_{*,n}, T_{1,n}) = \left(\frac{\mu_*}{\mu_1}\right)^2 = \left(\frac{u_*\sqrt{\frac{\pi(1-\pi)}{p_*^{(0)}q_*^{(0)}}}}{u_1\sqrt{\frac{\pi(1-\pi)}{p_1^{(0)}q_1^{(0)}}}}\right)^2 = \frac{u_*^2 p_1^{(0)} q_1^{(0)}}{u_1^2 p_*^{(0)} q_*^{(0)}} \qquad (3.14)$$

As mentioned in Section 3.4.2, in practice, we want to apply the ARE method for fixed alternatives. However, the ARE criterion given in (3.14) is based on alternative hypotheses close to 0. Our approach to define the ARE method for fixed alternatives is to use an approximation to the asymptotic value of ARE.

The ARE method for fixed alternatives, based on the two-sample score tests $T_{*,n}$ and $T_{1,n}$, is defined as:

$$are_p(p_1^{(0)}, p_2^{(0)}, p_1^{(1)}, p_2^{(1)}, \rho) = \left(\frac{p_*^{(0)} - p_*^{(1)}}{\sqrt{p_*^{(0)}q_*^{(0)}}}\right)^2 \Bigg/ \left(\frac{p_1^{(0)} - p_1^{(1)}}{\sqrt{p_1^{(0)}q_1^{(0)}}}\right)^2 \qquad (3.15)$$

The are_p approximates the ARE definition given in (3.14). Note that we have considered the fixed treatment effect stabilized for the sample size as an approximate value for the limiting treatment, that is: $\sqrt{n}(p_1^{(0)} - p_1^{(1)}) \cong u_1$ and $\sqrt{n}(p_*^{(0)} - p_*^{(1)}) \cong u_*$.

By using that the probability of the composite endpoint can be written by means of the probabilities of the composite components and the correlation (see (3.7)), the are is then in terms of the following anticipated parameters: (i) $p_1^{(0)}$ and $p_2^{(0)}$, event rates in control group for the relevant endpoint, ε_1, and the additional endpoint, ε_2; (ii) $p_1^{(0)} - p_1^{(1)}$ and $p_2^{(0)} - p_2^{(1)}$ fixed treatment effects for ε_1 and ε_2; (iii) $\rho^{(0)}$ and $\rho^{(1)}$, Pearson's correlation in each group.

3.7.2 Asymptotic invariance of ARE

In this section, we study the relationship between ARE and ARE_p given in (3.10) and (3.14), respectively. We will see that these two measures are, under certain

conditions, invariant. To do so, we first compare the speed of convergence of sequences of alternatives using differences of proportions over sequences of alternatives using odds ratios.

Convergence speed for contiguous alternatives

Let $p_{p,n}^{(1)}$ be the sequence of probabilities in the treatment group corresponding to contiguous alternatives of the form $H_n:\ p^{(0)} - p_{p,n}^{(1)} = \frac{u}{\sqrt{n}}$, where $u > 0$. Then, $p_{p,n}^{(1)}$ is given by:

$$p_{p,n}^{(1)} = p^{(0)} - \frac{u}{\sqrt{n}} \tag{3.16}$$

Let $p_{\beta,n}^{(1)}$ be the sequence of probabilities in the treatment group corresponding to contiguous alternatives of the form $H_n : \log(\text{OR})_n = \frac{v}{\sqrt{n}}$, where $v < 0$. Denoting by $\beta_n = \log(\text{OR})_n$, we have that:

$$\beta_n = \frac{v}{\sqrt{n}} \qquad \text{where} \qquad \beta_n = \log\left(\frac{p_{\beta,n}^{(1)}/q_{\beta,n}^{(1)}}{p^{(0)}/q^{(0)}}\right)$$

where $q_{\beta,n}^{(1)} = 1 - p_{\beta,n}^{(1)}$. Therefore, $p_{\beta,n}^{(1)}$ is given by:

$$p_{\beta,n}^{(1)} = e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}\left(1 + e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}\right)^{-1} \tag{3.17}$$

Note that both $p_{p,n}^{(1)}$ and $p_{\beta,n}^{(1)}$ converge to $p^{(0)}$ as $n \to +\infty$. In order to compare the speed of convergence of $p_{p,n}^{(1)}$ and $p_{\beta,n}^{(1)}$, we define the limiting ratio of their corresponding distances to the $p^{(0)}$ as:

$$L = \lim_{n \to +\infty} \frac{p_{p,n}^{(1)} - p^{(0)}}{p_{\beta,n}^{(1)} - p^{(0)}}$$

which has the following interpretation:

- If $L = 0$, the sequence $p_{p,n}^{(1)}$ converges faster than $p_{\beta,n}^{(1)}$.
- If $L = \infty$, the sequence $p_{\beta,n}^{(1)}$ converges faster than $p_{p,n}^{(1)}$.
- If $L = 1$, both sequences converge to null hypothesis at the same time, i.e., they have the same speed of convergence.

- If $L \in (0, +\infty)$, the rates of convergence are proportional.

In the following proposition, we state the expression of the limiting ratio L when considering the sequences $p^{(1)}_{p,n}$ and $p^{(1)}_{\beta,n}$ given in (3.16) and (3.17).

Proposition 3.1. *Let $p^{(0)}$ be the probability under the group 0, and let $p^{(1)}_{p,n}$ and $p^{(1)}_{\beta,n}$ be the sequences of probabilities given in* (3.16) *and* (3.17)*. Then:*

$$L = -\frac{u}{v}\frac{1}{p^{(0)}q^{(0)}} \in (0, +\infty)$$

Consequently, the convergence to the null hypothesis is proportional among parametrizations.

Proof. By considering $p^{(1)}_{p,n}$ and $p^{(1)}_{\beta,n}$ given in (3.16) and (3.17), then we have:

$$\begin{aligned}
L &= \lim_{n\to+\infty} \frac{p^{(1)}_{p,n} - p^{(0)}}{p^{(1)}_{\beta,n} - p^{(0)}} = \lim_{n\to+\infty} \frac{-\frac{u}{\sqrt{n}}}{\frac{e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}}{1+e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}} - p^{(0)}} = \lim_{n\to+\infty} \frac{-\frac{u}{\sqrt{n}}}{\frac{e^{\frac{v}{\sqrt{n}}}}{\frac{q^{(0)}}{p^{(0)}}+e^{\frac{v}{\sqrt{n}}}} - p^{(0)}} \\
&\overset{x=\frac{1}{\sqrt{n}}}{=} \lim_{x\to 0} \frac{-ux}{\frac{e^{vx}}{\frac{q^{(0)}}{p^{(0)}}+e^{vx}} - p^{(0)}} \overset{\text{H}}{=} \lim_{x\to 0} \frac{-u}{\frac{ve^{vx}(\frac{q^{(0)}}{p^{(0)}}+e^{vx}) - e^{vx}ve^{vx}}{\left(\frac{q^{(0)}}{p^{(0)}}+e^{vx}\right)^2}} = \lim_{x\to 0} \frac{-u}{\frac{ve^{vx}(\frac{q^{(0)}}{p^{(0)}})}{\left(\frac{q^{(0)}}{p^{(0)}}+e^{vx}\right)^2}} \\
&= -\frac{u}{v}\frac{1}{p^{(0)}q^{(0)}}
\end{aligned}$$

where H denotes the application of l'Hôpital rule. Finally, since $u > 0$, $v < 0$, and $p^{(0)}q^{(0)} > 0$, we have $L > 0$.

Asymptotic invariance

Based on the results from Section 3.7.2.1, we now focus on the invariance between the ARE in terms of the odds ratio (defined in (3.10)) and the ARE in terms of the risk difference (defined in (3.14)).

We start noticing that given the event rates $p^{(0)}_1$ and $p^{(0)}_*$ and the sequences of alternatives

$$\{p^{(1)}_{\beta,k}\}_n = e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}\left(1 + e^{\frac{v}{\sqrt{n}}}\frac{p^{(0)}}{q^{(0)}}\right)^{-1} \quad \text{and} \quad \{p^{(1)}_{p,k}\}_n = p^{(0)}_k - \frac{u_k}{\sqrt{n}},$$

it follows from the Proposition 3.1 that:

$$L_1 = -\frac{u_1}{v_1}\frac{1}{p_1^{(0)}q_1^{(0)}} \qquad \text{and} \qquad L_* = -\frac{u_*}{v_*}\frac{1}{p_*^{(0)}q_*^{(0)}}.$$

which are the convergence speed for contiguous alternatives using the risk difference over the odds ratio for the relevant and the composite endpoints, respectively.

Theorem 3.1. *Let $p_k^{(0)}$ be the event rate in the control group for the k-th endpoint ($k = 1, *$). Consider the hypotheses in terms of the odds ratio given in* (3.9) *and the hypotheses in terms of the risk difference given in* (3.13)*. Suppose $L_k \in (0, +\infty)$ defined by:*

$$L_k = -\frac{u_k}{v_k}\frac{1}{p_k^{(0)}q_k^{(0)}}, \qquad k = 1, *$$

Then, if $L_1 = L_*$, the ARE is invariant with respect to the parametrization, i.e.:

$$\mathrm{ARE}_p(T_*, T_1) = \mathrm{ARE}_\beta(T_*, T_1)$$

Proof. The proof is straightforward by applying Proposition 3.1 and then noticing that:

$$v_k^2 p_k^{(0)} q_k^{(0)} = \left(-\frac{1}{L}u_k\frac{1}{p_k^{(0)}q_k^{(0)}}\right)^2 p_k^{(0)}q_k^{(0)} = \left(\frac{1}{L}\right)^2 \frac{u_k^2}{p_k^{(0)}q_k^{(0)}}$$

where $L = L_1 = L_*$ and for $k = 1, *$. Hence, we obtain:

$$\mathrm{ARE}_\beta(T_*, T_1) = \frac{v_*^2 p_*^{(0)} q_*^{(0)}}{v_1^2 p_1^{(0)} q_1^{(0)}} = \frac{\left(\frac{1}{L}\right)^2 \frac{u_*^2}{p_*^{(0)}q_*^{(0)}}}{\left(\frac{1}{L}\right)^2 \frac{u_1^2}{p_1^{(0)}q_1^{(0)}}} = \frac{\frac{u_*^2}{p_*^{(0)}q_*^{(0)}}}{\frac{u_1^2}{p_1^{(0)}q_1^{(0)}}} = \mathrm{ARE}_p(T_*, T_1)$$

The assumption $L_1 = L_*$ may be thought of as representing that both endpoints have the same proportionality of the speed of convergence. However, further research is needed for the deep understanding and the implications of this condition.

3.8 Discussion

In this chapter, we have proposed a method that allows an informed selection between a binary composite endpoint or one of its components as primary endpoint.

Although composite endpoints are widely used as primary endpoints in clinical trials, as we have seen, they are not always the best option. The law governing the composite endpoint depends on the event rates, the magnitude of the treatment effects and the correlation between the components that form the composite. While the event rates and magnitude of the treatment effects can be reasonably well anticipated, this is not the case for the correlation between endpoints. Our methodology, and hence, the computation of the ARE has been established for different correlation values in each treatment group. However, the scenarios to derive the guidelines have been restricted to the same correlation in both groups. The impact of this assumption as well as the scenarios with two different correlations remain as future work.

If at least we could anticipate the degree of association in terms of weak, medium or large, we could use Table 3.4 to decide which endpoint to use. The treatment effects of the relevant and the additional endpoints also have an important role for deciding the primary endpoint. As seen earlier in Table 3.3, when the additional endpoint presents a smaller treatment effect than the relevant endpoint, it could not be more efficient to base the trial on the composite instead of the relevant endpoint, since the effect of the therapy in these settings could be diluted by adding an endpoint.

In order to assess the appropriate choice of the primary endpoint, we have created an interactive web-platform called CompARE to calculate the ARE method based on the information of the different endpoints together with anticipated values. We will present CompARE and describe its features in Chapter 5.

This work has been restricted to composite endpoints defined by two components. The method could be used for composite endpoints formed by more than two components by identifying two subsets of possible components (S_R and S_A) and then comparing the composite versus one of its subsets, for instance, S_R.

We have developed the ARE method for two different parametrizations (odds ratio and difference of proportions) and shown that the ARE is asymptotically invariant with respect to the parametrization in Section 3.7. Further work (not shown), however, suggest that this property is not longer true for fixed alternatives, that is to say, that are and are_p given in (3.11) and (3.15), respectively, are not equal. Further research is needed to explore whether or not this issue could affect the decision of using the composite endpoint or the relevant endpoint as the primary endpoint. Also, in addition to the odds ratio and risk difference parametrizations, we could have considered the hypothesis problem in terms of the risk ratio. The are method in terms of the risk ratio might be a topic of future research.

The standard definition of the Asymptotic Relative Efficiency relates the efficiency of two statistic tests for the same set of hypothesis. In this case, it can be interpreted as the limiting ratio of sample sizes to give the same asymptotic power under sequences of contiguous alternatives (Noether, 1954). Gómez and Gómez-Mateu (2014) empirically proved that the interpretation of the ARE as the ratio of required sample sizes still holds when using two logrank tests to compare the hazard ratios under the relevant or the composite endpoint. It remains to be seen whether the *are* we have proposed for binary endpoints can as well be interpreted as a ratio of required sample sizes.

Finally, in this work, the ARE method has been developed for discussing the use of a composite or one of its components as primary endpoint. We have assumed that both endpoints, ε_1 and ε_2, are important enough to be considered into the study and that one of the endpoints, ε_1, is more relevant than the other, ε_2. However, the ARE method does not take into account the relative relevance between ε_1 and ε_2. We understand this could be an important issue and remains open for future research.

Chapter 4

A class of statistics for binary and time-to-event endpoints

The work presented in this chapter reproduces the following paper currently under review:

A class of two-sample nonparametric statistics for binary and time-to-event outcomes.
Bofill Roig, M., and Gómez Melis, G.
arXiv:2002.01369 [stat.ME]

Appendix C contains the proof of theorems and the derivation and estimation of the covariance. The methodology presented in this chapter has been implemented in an R package called `Survbin`. We postpone the software's description to Chapter 5. The source code to reproduce the results of this work are online in the GitHub repository: `https://github.com/MartaBofillRoig/SurvBin`.

4.1 Introduction

In many clinical studies, two or more endpoints are investigated aiming to provide a comprehensive picture of the treatment's benefits and harms. Survival analysis has often been the sharp focus of clinical trial research. However, when there is more than one event of interest, the time until the appearance of the event is not always the unique center of attention; often the occurrence of an event over a fixed time period is as well an outcome of interest.

In the context of cancer immunotherapies trials, short-term binary endpoints based on the tumor size, such as objective response, are common in early-phase trials, whereas overall survival remains the gold standard in late-phase trials (Wilson et al., 2015; Ananthakrishnan and Menon, 2013). Since traditional oncology endpoints may not capture the clinical benefit of cancer immunotherapies, the idea of looking at both tumor response and survival has grown from the belief that together may achieve a better characterization of the clinical response (Thall, 2008).

Several authors have considered both objective response and overall survival as primary endpoints in cancer trials. Lai and ZeeLai and Zee (2015) proposed a single-arm phase II trial design with tumor response rate and a time-to-event outcome, such as overall survival or progression free survival. In their design, the dependence between the binary response and the time-to-event outcome is modeled through a Gaussian copula. Lai et al. (2012) proposed a two-step sequential design in which the response rate and the time to the event are jointly modeled. Their approach relates the response rate and the time to the event by means of a mixture model build on the basis of the Cox proportional hazards model assumption. Chen and Wang (2020) presented a joint model for binary marker responses and survival outcomes for clustered data. They based the statistical inference on a multivariate penalized likelihood method and estimate the standard errors using a jackknife resampling method.

An additional challenge in immunotherapy trials lies in the fact that delayed effects are likely to be found, bringing the need of alternative methods accounting for the non-proportionality of the hazards (Mick and Chen, 2015). Statistics that look at differences between integrated weighted survival curves, such as those defined by Pepe and Fleming (1989, 1991) and extended by Gu et al. (1999), are better suited to detect early or late survival differences and do not depend on the proportional hazards assumption. In this work, we aim to propose a class of two-sample statistics that could be used in seamless phase II/III design to jointly

evaluate the efficacy on binary and survival endpoints, even in the presence of delayed treatment effects.

The problem of how to analyze multiple outcomes has been widely discussed in the literature (Dmitrienko and Agostino, 2013; Alosh et al., 2014). The classical approach is to restrict the attention to multiple testing procedures that control the probability of one or more false rejections, the so-called familywise error rate, which guarantee the nominal significance level (Lehmann and Romano, 2005). However, classical multiple testing procedures based on correcting the significance level (e.g, Bonferroni procedure (Bland and Altman, 1995)) may not be appropriate since they do not take into account the potential association between the binary and survival outcomes and might lead to conservative designs.

Other alternative approaches have been developed allowing for the joint distribution of test statistics. O'Brien (1984) and Pocock et al. (1987) proposed global test statistics through the sum of individual statistics. O'Brien (1984) developed a generalized least squares method by combining multiple statistics into a single hypothesis test when variables are normally distributed; whereas Pocock et al. (1987) extended O'Brien's work to asymptotically normal test statistics. Hothorn et al. (2008) and Pipper et al. (2012) approached the problem of testing multiple hypothesis using parametric and semi-parametric models. Hothorn et al. (2008) used the limiting distribution of the parameter estimators to build upon the corresponding test statistics and their joint distribution. Based on that, their approach corrects the significance level by means of the simultaneous asymptotic normality of the test statistics. Pipper et al. (2012) proposed a procedure for evaluating the efficacy in trials with multiple endpoints of different types. Their procedure is based on simultaneous asymptotic normality of the effect estimators from the single-models for each endpoint together with multiple testing adjustments.

Extensive research has been done on joint modeling of longitudinal measurements and survival data (comprehensive overviews can be found in Tsiatis and Davidian (2004), Rizopoulos (2012) and Papageorgiou et al. (2019)). In most cases, the primary focus is on characterizing the association between the longitudinal and event time processes. The common framework is to relate the time-to-event and longitudinal outcomes through the proportional hazard model. Nevertheless, the relationship between binary response at a specific time point and survival outcome has received less attention (Chen and Wang, 2020).

In this work, we have followed the idea launched by Pocock et al. (1987) of combining multiple test statistics into a single hypothesis test. Specifically, we propose a class of statistics based on a weighted sum of a difference in proportions test and a weighted Kaplan-Meier test-based for the difference of survival

functions. Our proposal adds versatility into the study design by enabling different follow-up periods for each endpoint, and flexibility by incorporating weights. We define these weights to specify unequal priorities to the different endpoints and to anticipate the type of time-to-event difference to be detected.

This article is organized as follows. In Section 2, we present the class of statistics for binary and time-to-event outcomes. In Section 3, we set out the assumptions and present the large sample distribution theory for the proposed statistics. In Section 4, we introduce different weights and discuss their choice. We give an overview of our R package `SurvBin` in Section 5 and illustrate our proposal with a recent immunotherapy trial in Section 6. In Section 7, we evaluate the performance of these statistics in terms of the significance level and the statistical power with a simulation study. We conclude with a discussion.

All the required functions to use these statistics have been implemented in R and have been made available at: https://github.com/MartaBofillRoig/SurvBin.

4.2 A general class of binary and survival test statistics

Consider a study comparing two groups, control group ($i = 0$) and intervention group ($i = 1$), each composed of $n^{(i)}$ individuals, and denote by $n = n^{(0)} + n^{(1)}$ the total sample size. Suppose that both groups are followed over the time interval $[0, \tau]$ and are compared on the basis of the following two endpoints: the occurrence of an event ε_b before τ_b ($0 < \tau_b \leq \tau$), and the time to a different event ε_s within the interval $[\tau_0, \tau]$ ($0 \leq \tau_0 < \tau$). For the i-th group ($i = 0, 1$), let $p^{(i)}(\tau_b)$ be the probability of having the event ε_b before τ_b, and $S^{(i)}(\cdot)$ be the survival function of the time to the event ε_s.

We consider the problem of testing simultaneously $\mathrm{H}_{b,0}$: $p^{(0)}(\tau_b) = p^{(1)}(\tau_b)$ and $\mathrm{H}_{s,0}$: $S^{(0)}(t) = S^{(1)}(t), \forall t \in [\tau_0, \tau]$, aiming to demonstrate either a higher probability of the occurrence of ε_b or an improved survival with respect to ε_s in the intervention group. The hypothesis problem can then be formalized as:

$$\begin{cases} \mathrm{H}_0: & p^{(0)}(\tau_b) = p^{(1)}(\tau_b) \ \text{ and } \ S^{(0)}(t) = S^{(1)}(t), \ \forall t \in [\tau_0, \tau] \\ \mathrm{H}_1: & p^{(0)}(\tau_b) < p^{(1)}(\tau_b) \ \text{ or } \ S^{(0)}(t) \leq S^{(1)}(t), \ \ \forall t \in [\tau_0, \tau], \\ & \qquad\qquad \exists t^* \in [\tau_0, \tau], \ S^{(0)}(t^*) < S^{(1)}(t^*) \end{cases} \tag{4.1}$$

We propose a class of statistics –hereafter called $\mathcal{L}$-class– as a weighted linear combination of the difference of proportions statistic for the binary outcome and the integrated weighted difference of two survival functions for the time-to-event

outcome, as follows,

$$\mathbf{U}_n^{\omega}(\tau_0, \tau_b, \tau; \hat{Q}) = \omega_b \cdot \frac{U_{b,n}(\tau_b)}{\hat{\sigma}_b} + \omega_s \cdot \frac{U_{s,n}(\tau_0, \tau; \hat{Q})}{\hat{\sigma}_s} \tag{4.2}$$

for some real numbers $\omega_b, \omega_s \in (0,1)$, such that $\omega_b + \omega_s = 1$, and where:

$$U_{b,n}(\tau_b) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\hat{p}^{(1)}(\tau_b) - \hat{p}^{(0)}(\tau_b)\right) \tag{4.3}$$

$$U_{s,n}(\tau_0, \tau; \hat{Q}) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\int_{\tau_0}^{\tau} \hat{Q}(t) \cdot \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t)\right) dt\right) \tag{4.4}$$

denoting by $\hat{p}^{(i)}(\tau_b)$ the estimated proportion of events ε_b before τ_b, and by $\hat{S}^{(i)}(\cdot)$ the Kaplan-Meier estimator of $S^{(i)}(\cdot)$ for group i. The estimates $\hat{\sigma}_b^2$ and $\hat{\sigma}_s^2$ are such that converge in probability to σ_b^2 and σ_s^2, respectively, as $n \to +\infty$, where σ_b^2 and σ_s^2 represent the asymptotic variances of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$, respectively. Both theoretical and estimated expressions for the variances of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$ will be given in Section 4.3 (see equations (4.5, 4.6) for the theoretical expressions and (4.10, 4.11) for the estimates). The term $\hat{Q}(\cdot)$ is a possibly random function which converges pointwise in probability to a deterministic function $Q(\cdot)$. For ease of notation, and letting $\omega = (\omega_b, \omega_s)$, we will suppress the dependence on τ_0, τ_b, τ and use instead $\mathbf{U}_n^{\omega}(\hat{Q})$, $U_{b,n}$, $U_{s,n}(\hat{Q})$. Note that $\hat{p}^{(i)}(\tau_b)$, $\hat{S}^{(i)}(\cdot)$, $\hat{\sigma}_b$ and $\hat{\sigma}_s$ depend on the sample size $n^{(i)}$, but it has been omitted in notation for short.

The weights ω control the relative relevance of each outcome -if any- and the random weight function $\hat{Q}(\cdot)$ serves two purposes: to specify the type of survival differences that may exist between groups and to stabilize the variance of the difference of the two Kaplan-Meier functions. Some well-known special cases of $\hat{Q}(\cdot)$ are:

(i) $\hat{Q}(t) = \hat{G}(t-)$, where $\hat{G}(t-)$ is the pooled Kaplan-Meier estimator for the censoring distribution. This choice of $\hat{Q}(t)$ down-weights the contributions on those times where the censoring is heavy.

(ii) $\hat{Q}(t) = \hat{S}(t-)^{\rho} \cdot (1 - \hat{S}(t-))^{\gamma}$, where $\rho, \gamma \geq 0$ and $\hat{S}(t-)$ is the pooled Kaplan-Meier estimator for the survival function. This $\hat{Q}(t)$ corresponds to the weights of the Fleming-Harrington $G^{p,q}$ family (Fleming and Harrington, 1991). Then, for instance, if $\rho = 1$ and $\gamma = 0$, $\hat{Q}(t)$ emphasizes early differences between survival functions; whereas late differences could be highlighted with $\rho = 0$ and $\gamma = 1$.

(iii) $\hat{Q}(t) = \bar{Y}(t-)$, where $\bar{Y}(t-)$ denotes the number of individuals at risk of ε_s at time t. In this case $\hat{Q}(t)$ accentuates the information at the beginning of the survival curve allowing early failures to receive more weight than later failures.

We state the precise conditions for the weight function $\hat{Q}(\cdot)$ in Section 4.3 and postpone the discussion about the choice of $\hat{Q}(\cdot)$ and $\omega = (\omega_b, \omega_s)$ to Section 4.4.

The statistics in the $\mathcal{L}$-class are defined for possible different follow-up configurations based on different choices of: the overall follow-up period, τ; the time where the binary event is evaluated, τ_b; and the origin time for the survival outcome, τ_0; taking into account that $0 < \max\{\tau_0, \tau_b\} < \tau$. There are however no restrictions on whether or not these periods overlap and, if they do, how much and when. We illustrate two different situations with different configurations for τ_0, τ_b, τ in Figure 4.1. The first case is exemplified by an HIV therapeutic vaccination study where safety-tolerability response (binary outcome) and time-to-viral rebound (survival outcome) are outcomes of interest. Whereas the safety-tolerability is evaluated at week 6 ($\tau_b = 6$), the time-to-viral rebound is evaluated from week 6 to 18 ($\tau_0 = 6$ and $\tau = 18$) (De Jong et al., 2019). The second example in the area of immunotherapy trials includes a binary outcome (objective response), evaluated at month 6, and overall survival, evaluated from randomization until year 4 ($\tau_0 = 0$, $\tau_b = 0.5$ and $\tau = 4$) (Hodi et al., 2010).

The $\mathcal{L}$-class statistics includes several statistical tests. If $\tau_0 = 0$, $\tau_b = \tau$ and $\omega_b = \omega_s$, then, $\mathbf{U}_n^{\omega}(\hat{Q})$ corresponds to the global test statistic proposed by Pocock et al. (1987). If $\varepsilon_b = \varepsilon_s$, $\tau_0 = \tau_b$, and $\omega_b = \omega_s$, the statistic $\mathbf{U}_n^{\omega}(\hat{Q})$ is the equivalent of the linear combination test of Logan et al. (2008) when there is no censorship until τ_b for testing for differences in survival curves after a pre-specified time-point.

4.3 Large sample results

In this section, we derive the asymptotic distribution of the $\mathcal{L}$-class of statistics given in (4.2) under the null hypothesis and under contiguous alternatives, present an estimator of their asymptotic variance, and discuss the consistency of the $\mathcal{L}$-statistics against any alternative hypothesis of the form of H_1 in (4.1). We start the section with the conditions we require for the $\mathcal{L}$-class of statistics. In order to make the work more concise and more readable, proofs and technical details are in Appendix C.

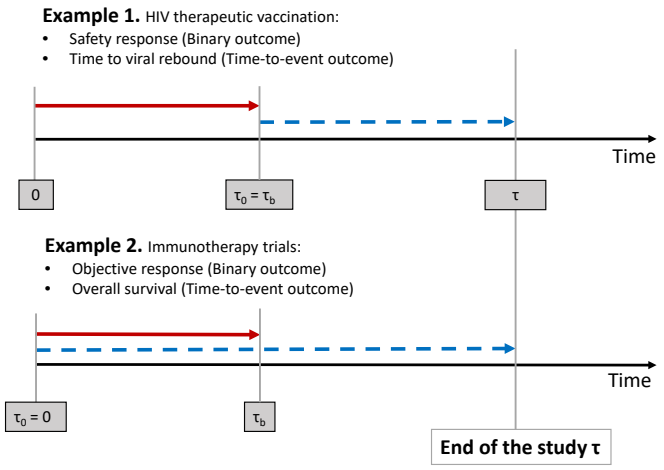


Fig. 4.1 Illustration of two different follow-up configurations, the red and blue arrows represent the time-frame for binary and time-to-event outcomes, respectively. The red line goes from the start of the study (at time-point 0) until the binary outcome is evaluated at time τ_b. The blue (dashed) line goes from when the time-to-event information begins to be collected (τ_0) to the end of the study (τ).

4.3.1 Further notation and Assumptions

We consider two independent random samples of $n^{(i)}$ $(i = 0, 1)$ individuals and for each we denote the binary response by $X_{ij} = I\{\varepsilon_b \text{ has occurred}\}$, the time to ε_s by T_{ij} and the censoring time by C_{ij} for $j = 1, ..., n^{(i)}$ and where $I\{\cdot\}$ is the usual 0/1 indicator function. Assume that T_{ij} is non-informatively right-censored by C_{ij}, that X_{ij} is independent of C_{ij}, and that the occurrence of the survival and censoring times, T_{ij} and C_{ij}, does not prevent to assess the binary response, X_{ij}. The observable data are summarized by $\{X_{ij}, T_{ij} \wedge C_{ij}, \delta_{ij}\}$, where $\delta_{ij} = I\{T_{ij} \wedge C_{ij} = T_{ij}\}$ and $a \wedge b = \min(a, b)$.

Denote by $G^{(i)}(\cdot)$ and $\hat{G}^{(i)}(\cdot)$ the censoring survival function and the Kaplan-Meier estimator for the censoring times, respectively. As we will see in the next section, the distribution of the $\mathcal{L}$-statistics relies, among others, on the survival function for those individuals who respond $(X_{ij} = 1)$ to the binary endpoint, We then introduce here the survival function for responders as $S_X^{(i)}(t) = \mathrm{P}(T_{ij} > t | X_{ij} = 1)$ $(t > \tau_b)$.

Furthermore we assume that: (i) at the end of follow-up, $S^{(i)}(\tau) > 0$, $S_X^{(i)}(\tau) > 0$ and $G^{(i)}(\tau) > 0$; (ii) the limiting fraction of the total sample size is non-negligible,

i.e., $\lim_{n\to+\infty} n^{(i)}/n = \pi^{(i)} \in (0,1)$; and (iii) $Q(\cdot)$ is a nonnegative piecewise continuous with finitely discontinuity points. For all the continuity points in $[0,\tau]$, $\hat{Q}(t)$ converges in probability to $Q(t)$ as $n \to +\infty$. Moreover, $\hat{Q}(\cdot)$ and $Q(\cdot)$ are functions of total variation bounded in probability.

Finally, we introduce the counting process $\bar{N}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} N_{ij}(t) = \sum_{j=1}^{n^{(i)}} I\{T_{ij} \wedge C_{ij} \leq t, \delta_{ij} = 1\}$ as the number of observed events that have occurred by time t for the i-th group $(i = 0, 1)$ and $\bar{Y}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} Y_{ij}(t) = \sum_{j=1}^{n^{(i)}} I\{T_{ij} \wedge C_{ij} \geq t\}$ as the number of subjects at risk at time t for the i-th group. We define $y^{(i)}(s) = \mathrm{E}\left(Y_{ij}(s)\right)$ and suppose that $y^{(i)}(\tau) > 0$.

4.3.2 Asymptotic distribution

In order to derive the asymptotic distribution of the statistic $\mathbf{U}_n^{\omega}(\hat{Q})$, we use that $\mathbf{U}_n^{\omega}(\hat{Q})$ can be approximated by $\mathbf{U}_n^{\omega}(Q)$, the same statistic with the weights replaced by its deterministic function (see Lemma 1 in Appendix C). Roughly speaking, thanks to this approximation we can ignore the randomness of $\hat{Q}(\cdot)$ and use $\mathbf{U}_n^{\omega}(Q)$ to obtain the limiting distribution of $\mathbf{U}_n^{\omega}(\hat{Q})$. In what follows, we state the asymptotic distributions under the null hypothesis in Theorem 4.1 and under a sequence of contiguous alternatives in Theorem 4.2.

Theorem 4.1. *Let $\boldsymbol{U}_n^{\omega}(\hat{Q})$ be the statistic defined in* (4.2)*. Under the conditions outlined in 4.3.1, if the null hypothesis $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ holds, $\boldsymbol{U}_n^{\omega}(\hat{Q})$ converges in distribution, as $n \to +\infty$, to a normal distribution as follows:*

$$\boldsymbol{U}_n^{\omega}(\hat{Q}) \to N\left(0, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \cdot \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

where σ_b^2, σ_s^2 stand for the asymptotic variances of $U_{b,n}$ and $U_{s,n}(Q)$, respectively, and σ_{bs} is the covariance between $U_{b,n}$ and $U_{s,n}(Q)$. Their corresponding expressions are given by:

$$\sigma_b^2 = \sum_{i=0,1} \left(1-\pi^{(i)}\right) p^{(i)}\left(\tau_b\right)\left(1-p^{(i)}\left(\tau_b\right)\right) \tag{4.5}$$

$$\sigma_s^2 = -\sum_{i=0,1} \left(1-\pi^{(i)}\right) \int_{\tau_0}^{\tau} \frac{(K_\tau^{(i)}(t))^2}{(S^{(i)}(t))^2 G^{(i)}(t)} dS^{(i)}(t) \tag{4.6}$$

$$\sigma_{bs} = \sum_{i=0,1} \left(1-\pi^{(i)}\right) \cdot \left(I\{\tau_{\max} = \tau_b\} \cdot \int_{\tau_0}^{\tau_b} \frac{K_{\tau_b}^{(i)}(t)}{S^{(i)}(t)} \cdot \left(p_N^{(i)}(t) - p^{(i)}\left(\tau_b\right)\right) dS^{(i)}(t) \right.$$
$$\left. + \int_{\tau_{\max}}^{\tau} \frac{K_\tau^{(i)}(t)}{S^{(i)}(t)} \cdot p^{(i)}\left(\tau_b\right) \left(dS_X^{(i)}(t) - dS^{(i)}(t)\right) \right) \tag{4.7}$$

where $\tau_{\max} = \max(\tau_0, \tau_b)$, $K_{\tau_*}^{(i)}(t) = \int_t^{\tau_*} Q(u) S^{(i)}(u) du$ *(*$\tau_* = \tau$ *or* τ_b*),* $p_N^{(i)}(t) = P(X_{ij} = 1 | dN_{ij}(t) = 1)$*, and* $S_X^{(i)}(t) = P(T_{ij} > t | X_{ij} = 1)$ *for* $i = 0, 1$.

Recall that σ_b^2, σ_s^2, and σ_{bs} depend on τ_0, τ_b, τ, but we omit them for notational simplicity.

Theorem 4.2. *Let* $\boldsymbol{U}_n^\omega(\hat{Q})$ *be the statistic defined in* (4.2)*. Under the conditions outlined in 4.3.1, consider the following sequences of contiguous alternatives for both binary and time-to-event hypotheses satisfying, as* $n \to +\infty$*:*

$$\sqrt{n}(p_n^{(1)} - p^{(0)}) \to g$$

and

$$\sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) \to \mathcal{G}(t)$$

for some constant $g \in \mathbb{R}^+$ *and bounded function* $\mathcal{G}(\cdot) \in \mathbb{R}^+$*, and* $\forall t \in [\tau_0, \tau]$*. Then, under contiguous alternatives of the form:*

$$\mathrm{H}_{1,n} : \sqrt{n}(p_n^{(1)} - p^{(0)}) = g \;\; \textit{and} \;\; \sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) = \mathcal{G}(t), \quad \forall t \in [\tau_0, \tau] \tag{4.8}$$

we have that:

$$\boldsymbol{U}_n^\omega(\hat{Q}) \to N\left(\omega_b g + \omega_s \int_{\tau_0}^{\tau} Q(t)\mathcal{G}(t) dt, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

in distribution as $n \to +\infty$*, where* σ_b^2*,* σ_s^2 *and* σ_{bs} *are given in* (4.5)*,* (4.6) *and* (4.7)*, respectively.*

The covariance in (4.7) involves the conditional probabilities $S_X^{(i)}(t)$ and $p_N^{(i)}(t)$, while $S_X^{(i)}(t)$ represents the survival function for responders –individuals that have

had the binary event ε_b–, $p_N^{(i)}(t)$ stands for the probability of being a responder among individuals experiencing ε_s at t. Also note that, if $\tau_b < \tau_0$, the survival experience starts after the binary event has been evaluated and only involves the second integral in (4.7).

We notice that the efficiency of the $\mathcal{L}$-statistics, $\mathbf{U}_n^{\omega}(\hat{Q})$, under contiguous alternatives is driven by the non-centrality parameter $\mu_c = \omega_b g + \omega_s \int_{\tau_0}^{\tau} Q(t)\mathcal{G}(t)dt$, that is, by the sum of the weighted non-centrality parameters of $U_{b,n}$ and $U_{s,n}(\hat{Q})$.

4.3.3 Variance estimation and consistency

We now describe how to use the $\mathcal{L}$-statistics to test H_0 versus H_1 given in (4.1). In particular, we propose a consistent estimator of the asymptotic variance of $\mathbf{U}_n^{\omega}(\hat{Q})$, and present the standardized $\mathcal{L}$-statistics to test $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$.

The asymptotic variance of $\mathbf{U}_n^{\omega}(\hat{Q})$, given in Theorem 4.1, can be consistently estimated by:

$$\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q})) = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\hat{\sigma}_{bs}}{\hat{\sigma}_b \cdot \hat{\sigma}_s} \tag{4.9}$$

where $\hat{\sigma}_b$, $\hat{\sigma}_s$, and $\hat{\sigma}_{bs}$ denote the estimates of σ_b, σ_s and σ_{bs}, and are given by:

$$\hat{\sigma}_b^2 = \hat{p}\,(\tau_b)\,(1 - \hat{p}\,(\tau_b)) \tag{4.10}$$

$$\hat{\sigma}_s^2 = -\int_{\tau_0}^{\tau} \frac{(\hat{K}_\tau(t))^2}{\hat{S}(t)\hat{S}(t-)} \cdot \frac{n^{(0)}\hat{G}^{(0)}(t-) + n^{(1)}\hat{G}^{(1)}(t-)}{\hat{G}^{(0)}(t-)\hat{G}^{(1)}(t-)} d\hat{S}(t) \tag{4.11}$$

$$\begin{aligned}\hat{\sigma}_{bs} = &-\int_{\tau_0}^{\tau_b} \hat{K}_{\tau_b}(t) \left(\sum_{i=0,1} \frac{n - n^{(i)}}{n} \cdot \hat{\lambda}_{X,T}^{(i)}(t)dt + \frac{\hat{p}\,(\tau_b) \cdot d\hat{S}(t)}{\hat{S}(t)} \right) \\ &+ \int_{\tau_b}^{\tau} \frac{\hat{K}_\tau(t) \cdot \hat{p}\,(\tau_b)}{\hat{S}(t-)} \left(-\frac{\hat{S}(t-) \cdot d\hat{S}(t)}{\hat{S}(t)} + \sum_{i=0,1} \frac{n - n^{(i)}}{n} \cdot \frac{\hat{S}_X^{(i)}(t-) \cdot d\hat{S}_X^{(i)}(t)}{\hat{S}_X^{(i)}(t)} \right)\end{aligned} \tag{4.12}$$

where $\hat{K}_{\tau_*}(t) = \int_t^{\tau_*} \hat{Q}(u)\hat{S}(u)du$ ($\tau_* = \tau$ or τ_b), $\hat{S}(t)$ is the pooled Kaplan-Meier estimator of the survival functions, $\hat{p}\,(\tau_b)$ is the pooled estimator of the probabilities $p^{(i)}\,(\tau_b)$, $\hat{S}_X^{(i)}(t)$ is the Kaplan-Meier estimator of $S_X^{(i)}(t)$, and $\hat{\lambda}_{X,T}^{(i)}(t)$ is the estimator of $\lambda_{X,T}^{(i)}(t) = \lim_{dt \to 0} P(X_{ij} = 1, t \leq T_{ij} < t + dt | T_{ij} > t)/dt$.

The variance estimator presented in (4.9) is obtained assuming that the variances of the two groups are equal (pooled estimator). An unpooled variance estimator is proposed in Appendix C. For both pooled and unpooled estimators,

smoothing techniques are used to estimate $\lambda_{X,T}^{(i)}(t)$ over the time period $[\tau_0, \tau_b]$. In this work, we have chosen kernel smoothing methods. Note that resampling methods can also be used to get an estimator of the variance of $\mathbf{U}_n^{\omega}(\hat{Q})$. In the simulation section, we will discuss the results using the pooled, unpooled and bootstrap variance estimators.

In order to test the global null hypothesis $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ in (4.1), we consider the normalized statistic of $\mathbf{U}_n^{\omega}(\hat{Q})$:

$$\mathbf{U}_n^{\omega}(\hat{Q}) \Big/ \sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))} \tag{4.13}$$

Because this statistic (C.4) converges in distribution to a standard normal distribution, it can be used to test $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ by comparing $\mathbf{U}_n^{\omega}(\hat{Q}) \big/ \sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$ to a standard normal distribution. Moreover, for positive $Q(\cdot)$, the statistic is consistent against any alternative hypothesis of the form of H_1 in (4.1).

4.4 On the choice of weights

An important consideration when applying the statistics proposed in this work is the choice of the weight functions. The $\mathcal{L}$-class of statistics involves the already mentioned random weight function $\hat{Q}(t)$ and deterministic weights $\omega = (\omega_b, \omega_s)$. These weights are defined according to different purposes and have different roles into the statistic $\mathbf{U}_n^{\omega}(Q)$. In this section, we include different weights and discuss some of their strengths as well as shortcomings. The list provided is not exhaustive, other weights are possible and might be useful in specific circumstances.

4.4.1 Choice of $\omega = (\omega_b, \omega_s)$

The purpose of the weights ω is to prioritize the binary and the time-to-event outcomes. They have to be specified in advance according to the research questions. Whenever the two outcomes are equally relevant, we should choose $\omega_b = \omega_s = 0.5$. In this case the statistics will be optimal whenever the standardized effects on both outcomes coincide.

4.4.2 Choice of $\hat{Q}(\cdot)$

The choice of $\hat{Q}(\cdot)$ might be very general as long as $\hat{Q}(\cdot)$ converges in probability to a function $Q(\cdot)$, and both $\hat{Q}(\cdot)$ and $Q(\cdot)$ satisfy the conditions outlined in 4.3.1. In this section, we center our attention on a family of $\hat{Q}(\cdot)$ weights of the form:

$$\hat{Q}(t) = \hat{f}(t) \cdot \hat{v}(t),$$

where: (i) $\hat{f}(\cdot)$ is a data-dependent function that converges, in probability to $f(\cdot)$, a nonnegative piecewise continuous function with bounded variation on $[0, 1]$. The term $\hat{f}(t)$ takes care of the expected differences between survival functions and can be used as well to emphasize some parts of the follow-up according to the time-points (τ_0, τ_b, τ_s); (ii) the weights $\hat{v}(\cdot)$ converge in probability to a deterministic positive bounded weight function $v(\cdot)$. The main purpose of the weight $\hat{v}(t)$ is to ensure the stability of the variance of the difference of the two Kaplan-Meier functions. To do so, we make the additional assumption that:

$$|v(t)| \leq \Gamma \cdot G^{(i)}(t)^{1/2+\delta} \quad \text{and} \quad |\hat{v}(t)| \leq \Gamma \cdot \hat{G}^{(i)}(t)^{1/2+\delta}$$

for all $t \in [\tau_0, \tau]$, $i = 0, 1$ and for some constants $\Gamma, \delta > 0$.

Different choices of $\hat{f}(t)$ yield other known statistics. For instance, if $f(\cdot) = 1$, $U_{s,n}(\hat{Q})$ corresponds to the Weighted Kaplan-Meier statistics (Pepe and Fleming, 1989, 1991). Whenever $\hat{f}$ and $\hat{v}$ correspond to the weights (4.15) and (4.14), respectively, introduced below, we have the statistic proposed by Shen and Cai (2001). Furthermore, note that the weight functions of the form $\hat{Q}(t) = \hat{f}(t) \cdot \hat{v}(t)$ are similar to those proposed by Shen and Cai (2001); while they assume that $\hat{f}$ is a bounded continuous function, we assume that $\hat{f}(\cdot)$ is a nonnegative piecewise continuous function with bounded variation on $[0, 1]$, and instead of only considering the Pepe and Fleming weight function corresponding to (4.15), we also allow for different weight functions $\hat{v}(t)$. Finally, if we do not consider any weight, that is, if $\hat{Q}(t) = 1, \forall t$, $U_{s,n}(\hat{Q})$ corresponds to the difference of restricted mean survival times from τ_0 to τ.

In what follows, we outline different choices of $\hat{v}(t)$ and $\hat{f}(t)$, together with a brief discussion for each one:

- We require $\hat{v}(t)$ to be small towards the end of the observation period if censoring is heavy. The usual weight functions $\hat{v}(t)$ involve Kaplan-Meier estimators of the censoring survival functions. The most common weight functions are:

$$\hat{v}_c(t) = \frac{n\hat{G}^{(0)}(t-)\hat{G}^{(1)}(t-)}{n^{(0)}\hat{G}^{(0)}(t-) + n^{(1)}\hat{G}^{(1)}(t-)} \tag{4.14}$$

and $\hat{v}_{\surd}(t) = \sqrt{\hat{v}_c(t)}$, both proposed by Pepe and Fleming. Among other properties, $\hat{v}_c(\cdot)$ has been proved to be a competitor to the logrank test for the proportional hazards alternative (Pepe and Fleming, 1989). Note that if the censoring survival functions are equal for both groups and the sampling design is balanced ($n^{(0)} = n^{(1)}$), then, the differences in Kaplan-Meier estimators are weighted by the censoring survival function, that is, $w(t) = C(t) = C^{(i)}(t)$ for $i = 0, 1$. Also note that $w(t) = 1$ for uncensored data.

- Analogously to Fleming and Harrington(Fleming and Harrington, 1991) statistics, $\hat{f}(t)$ could be used to specify the type of expected differences between survival functions. That is, if we set:

$$f(\hat{S}(t-)) = \hat{S}(t-)^{\rho}(1 - \hat{S}(t-))^{\gamma}, \quad \rho, \gamma \geq 0 \tag{4.15}$$

  the choice $\rho > 0$, $\gamma = 0$ leads to a test to detect early differences, while $\rho = 0$, $\gamma > 0$ leads to a test to detect late differences; and $\rho = \gamma = 0$ leads to a test evenly distributed over time and corresponds to the weight function of the logrank.
- In order to put more emphasis on those times after the binary follow-up period we might consider:

$$f(t) = \begin{cases} a, & t < \tau_b \\ 1 - a, & t \geq \tau_b \end{cases}$$

  for $a < 0.5$.

4.5 Implementation

We have developed the `SurvBin` package to facilitate the use of the $\mathcal{L}$-statistics and is now available on GitHub (`https://github.com/MartaBofillRoig/SurvBin`). The `SurvBin` package contains two key functions: `lstats` to compute the standardized $\mathcal{L}$-statistic, $\mathbf{U}_n^{\omega}(\hat{Q})/\sqrt{\widehat{\text{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$, using the variance estimator given in Section 4.3; and `lstats_boots` to compute the standardized $\mathcal{L}$-statistic by using a bootstrap procedure to estimate the variance.

The `SurvBin` package also provides the functions `survbinCov` to calculate $\hat{\sigma}_{bs}$; and `bintest` and `survtest` to compute the univariate binary and survival statistics (4.3) and (4.4), $U_{b,n}(\tau_b)/\hat{\sigma}_b$ and $U_{s,n}(\tau_0, \tau; \hat{Q})/\hat{\sigma}_s$, respectively. In addition, the `SurvBin` package includes the function `simsurvbin` that can be used to simulate bivariate binary and survival data in a variety of situations.

The main function `lstats` can be called by:

```
lstats(time, status, binary, treat,
       tau0, tau, taub, rho, gam, eta, wb, ws, var_est)
```

where `time, status, binary` and `treat` are vectors of the right-censored data, the status indicator, the binary data and the treatment group indicator, respectively; `tau0, tau, taub` denote the follow-up configuration; `wb, ws` are the weights ω; `rho, gam, eta` are scalar parameters that controls the weight $\hat{Q}(t)$ which is given by $\hat{Q}(t) = \hat{G}(t-)^{\eta} \cdot \hat{S}(t-)^{\rho} \cdot (1 - \hat{S}(t-))^{\gamma}$; and `var_est` indicates the variance estimate to use (`pooled` or `unpooled`).

In this work, we estimate $\lambda^{(i)}_{X,T}(t)$ by means of the Epanechnikov kernel function, and the local bandwidth selection and the boundary correction described by Muller and Wang (1994) by using the `muhaz` package (Hess and Gentleman, 2019).

4.6 Example

Melanoma has been considered a good target for immunotherapy and its treatment has been a key goal in recent years. Here we consider a randomized, double-blind, phase III trial whose primary objective was to determine the safety and efficacy of the combination of a melanoma immunotherapy (gp100) together with an antibody vaccine (ipilimumab) in patients with previously treated metastatic melanoma (Hodi et al., 2010). Despite the original endpoint was objective response rate at week 12, it was amended to overall survival and then considered secondary endpoint. A total of 676 patients were randomly assigned to receive ipilimumab plus gp100, ipilimumab alone, or gp100 alone. The study was designed to have at least 90% power to detect a difference in overall survival between the ipilimumab-plus-gp100 and gp100-alone groups at a two-sided α level of 0.05, using a log-rank test. Cox proportional-hazards models were used to estimate hazard ratios and to test their significance. The results showed that ipilimumab with gp100 improved overall survival as compared with gp100 alone in patients with metastatic melanoma. However, the treatment had a delayed effect and an overlap between the Kaplan-Meier curves was observed during the first six months. Hence, the proportional hazards assumption appeared to be no longer valid, and a different approach would have been advisable.

In order to illustrate our proposal, we consider the comparison between the ipilimumab-plus-gp100 and gp100-alone groups based on the overall survival and

objective response as multiple primary endpoints of the study. For this purpose, we have reconstructed individual observed times by scanning the overall survival Kaplan-Meier curves reported in Figure 1A of (Hodi et al., 2010) using the `reconstructKM` package (Sun, 2020) (see Figure 4.2), and, afterwards, we have simulated the binary response to mimic the percentage of responses obtained in the study.

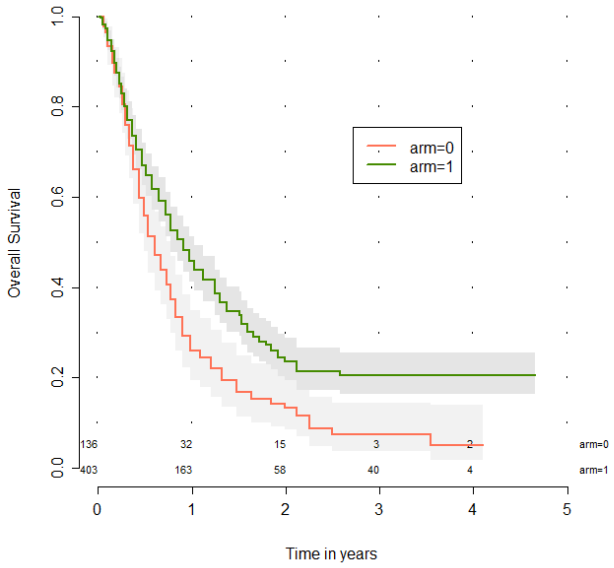


Fig. 4.2 Kaplan-Meier Curves for Overall Survival for ipilimumab-plus-gp100 and gp100-alone groups (arms 1 and 0, respectively).

Using the data obtained, we employ the $\mathcal{L}$-statistic by means of the function `lstats` in the `SurvBin` package. To do so, we need to specify the weights $(\hat{Q}, \omega)$ to be used, and the time-points (τ_0, τ_b, τ). In our particular case, we take $\tau_0 = 0, \tau_b = 0.5, \tau = 4$ according to the trial design, choose $\hat{Q}(t) = \hat{G}(t-) \cdot (1 - \hat{S}(t-))$ to account for censoring and delayed effects in late times, and $(\omega_b, \omega_s) = (0.25, 0.75)$ to emphasize the importance of overall survival over objective response.

As shown below, the function `lstats` returns the standardized $\mathcal{L}$-statistic, together with the $\mathcal{L}$-statistic and its standard deviation, and the individual statistics.

```
lstats(time=data$time, status=data$status, binary=data$
   binary, treat=data$treat,
       tau0=0, tau=4, taub=0.5, rho=0, gam=1, eta=1,
       wb=0.25, ws=0.75, var_est = "Pooled")
##
## $LTest
## Parameter                Value
## 1 (Standardized) L-Test 4.0950273
## 2                L-Test 3.2362929
## 3    Standard deviation 0.7902982
##
## $Binary_Tests
## Parameter                 Value
## Test Standardized L-Test 1.8678088
## Ub           Binary Test 0.4540763
## sd    Standard deviation 0.2431064
##
## $Survival_Tests
## Parameter                Value
## Test  Standardized Test 3.6924543
## Us         Survival Test 2.4398019
## sd   Standard deviation 0.6607534
##
## $Covariance
## Parameter     Value
## 1 Covariance -0.0001836297
```

The value of the $\mathcal{L}$-statistic, $\mathbf{U}_n^{\omega}(\hat{Q})$ in (4.2), is 3.24 and is obtained by using the values of $U_{b,n}(\tau_b)$ and $U_{s,n}(\tau_0, \tau; \hat{Q})$ (0.45 and 2.44, respectively), and $\hat{\sigma}_b$ and $\hat{\sigma}_s$ (0.24 and 0.66). The statistic $\mathbf{U}_n^{\omega}(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$ equals 4.10 and is computed by using the variance estimator in (4.9) and then by means of $\hat{\sigma}_b$ and $\hat{\sigma}_{bs}$ together with the estimated covariance $\hat{\sigma}_{bs}$ (-0.0002).

Since we obtained $\mathbf{U}_n^{\omega}(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))} = 4.10 > z_{\alpha=0.05}$, we have a basis to reject H_0 and conclude that the ipilimumab either improved overall survival or increased the percentages of tumor reduction in patients with metastatic melanoma, or both. Note that, in this example, we have been using the pooled variance estimator for the $\mathcal{L}$-statistic. We have also calculated the statistic using the unpooled and bootstrap variance estimators (see Appendix C) and notice that the results were not substantially different.

4.7 Simulation study

4.7.1 Design

We conducted a simulation study to evaluate our proposal in terms of the statistical power and the type-I error with small sample sizes. We generated bivariate binary and time-to-event data through a copula-based framework and used conditional sampling as described in Trivedi and Zimmer (2007).

The parameters used for the simulation (summarized in Table 4.1) have been the following:

- Frank's copula with association parameter between the marginal distributions of the binary and time-to-event outcomes equal to $\theta = 0.001, 2, 3$. These values correspond, respectively, to Spearman's rank correlation values equal to 0.0002, 0.32, 0.45 which represent increasing associations between the binary and time-to-event outcomes. We have not considered higher values of θ as they do not fulfill the condition that $S_X^{(i)}(\tau) > 0$ $(i = 0, 1)$.
- Weibull survival functions, $S_{b,a}^{(0)}(t) = e^{-(t/b)^a}$, with $a = 0.5, 1, 2$ and $b = 1$.
- Probability of having the binary endpoint $p^{(0)} = 0.1, 0.3$; and
- Sample size per arm $n^{(i)} = 250$.
- The censoring distributions between groups were assumed equal and uniform $U(0, c)$ with $c = 3$.
- Two different follow-up configurations were considered for $\tau_0 < \tau_b \leq \tau$: (i) $\tau_0 = 0, \tau_b = 0.5, \tau = 1$; and (ii) $\tau_0 = 0, \tau_b = \tau = 1$.
- We have considered the weights: $\hat{Q}(t) = \hat{G}(t-)^\eta \cdot \hat{S}(t-)^\rho \cdot (1 - \hat{S}(t-))^\gamma$ with $\eta = 1$ and $\rho, \gamma = 0, 1$. When simulating under the null hypothesis, we considered (ω_b, ω_s) equal to $(0.5, 0.5)$; whereas when simulating under the alternative hypotheses, we considered (ω_b, ω_s) equal to $(0.25, 0.75)$, $(0.5, 0.5)$, and $(0.75, 0.25)$.

The simulations under the alternative hypothesis considered four different situations depending on whether there is treatment effect on both endpoints and the type of difference between the survival curves. Specifically, the following cases were considered:

(1) Effect on both binary and survival endpoints. The effect on the survival endpoint satisfies the proportional hazards assumption, that is, the hazard ratio (HR) between treatment groups is constant over the study duration;
(2) Effect on the binary endpoint and non-effect on the survival endpoint ($H_{s,0}$: $S^{(0)}(t) = S^{(1)}(t),\ \forall t$);

(3) Non-effect on the binary endpoint ($H_{b,0} : p^{(0)}(\tau_b) = p^{(1)}(\tau_b)$) and effect on the survival endpoint with HR constant over the study duration;
(4) Effect on both binary and survival endpoints. The treatment differences on the survival endpoint have a delayed effect, that is, the survival functions are assumed to be equal until time t_*, and there is a constant hazard ratio (HR) between treatment groups from t_* to τ.

We used $d = p^{(1)}(\tau_b) - p^{(0)}(\tau_b) = 0.075$ to simulate the effects on the binary endpoint. For the survival endpoint, we considered HR$= 0.75$ under proportional hazards, and HR$= 0.70$ and $t_* = 0.5$ under delayed effects.

We evaluated the empirical significance level and the statistical power using the $\mathcal{L}$-statistics with pooled, unpooled and bootstrap variance estimators, and for the sake of the comparison, using the Bonferroni procedure. In addition, we presented the empirical results for testing the individual hypothesis $H_{b,0}$ and $H_{s,0}$ by using the statistics (4.3) and (4.4).

The total number of scenarios was 1456 (144 under the null hypothesis and 1312 under the alternative hypothesis). We ran 1000 replicates and estimated the significance level ($\alpha = 0.05$) for each scenario under the null hypothesis. We ran 100 replicates and estimated the statistical power for each scenario under alternative hypotheses. We performed all computations using the R software (version 4.0.2). The time required to perform the considered simulations was 89 hours.

Table 4.1 Scenarios used in the simulation study.

Parameter	Value	Parameter	Value
$p^{(0)}$	$0.1, 0.3$	a	$0.5, 1, 2$
b	1	c	3
θ	$0.001, 2, 3$	$n^{(i)}$ $(i = 0, 1)$	250
τ_b	$0.5, 1$	τ	1
ρ, γ	$0, 1$	η	1
d	$0, 0.075$	HR	$0.75, 1$
(ω_b, ω_s)	$(0.25, 0.75)$,$(0.5, 0.5)$,$(0.75, 0.25)$	t_*	$0, 0.5$

4.7.2 Power properties

When there is treatment effect on both endpoints and the proportional hazards assumption is fulfilled (case 1), we obtained empirical powers with medians 0.84, 0.84, 0.83 using the $\mathcal{L}$-statistics with pooled, unpooled and bootstrap variance estimators, respectively; whereas the median of the empirical powers using Bonferroni was 0.78. When there is treatment effect on both endpoints and there are delayed effects (case 4), the empirical powers for the $\mathcal{L}$-statistics have medians 0.79, 0.79, 0.77 using the pooled, unpooled and bootstrap variance estimators, respectively; whereas the median of the empirical powers using Bonferroni was 0.67.

Table 4.2 summarizes the simulation results on the power across different parameters in case 1. We compared the performance of the different variance estimators and noticed that the empirical powers do not substantially differ between them. We also observed that the power is not affected by the different weight functions in the case of proportional hazards. We obtained higher powers when emphasizing late-differences between the survival curves ($\gamma = 1$) in the case of delayed effects (median powers of 0.84, 0.83, 0.78 for unpooled, pooled and bootstrap variance estimators, respectively, with $\gamma = 1$ against 0.77, 0.74, 0.71 for unpooled, pooled and bootstrap variance estimators with $\gamma = 0$).

Figure 4.3 shows boxplots for the empirical powers using the pooled, unpooled and bootstrap variance estimators and using the Bonferroni procedure. These simulations show the superiority of the $\mathcal{L}$-statistics over the Bonferroni procedure, in terms of power, both under proportional hazards and under delayed effects and regardless of the choice of the weights (ω_b, ω_s).

When there is treatment effect only one of the endpoints (cases 2 and 3), the behavior of the power mainly relies on the pre-specified weights (ω_b, ω_s) (see Figure 4.3). If the survival endpoint is considered clinically more important than the binary endpoint and we use the weights $(\omega_b = 0.25, \omega_s = 0.75)$, then the median of the empirical powers is around 0.60 in case 2 (i.e., when there is treatment effect on the survival endpoint) and around 0.16 in case 3 (i.e., when there is no effect on the survival endpoint). We found a similar behavior when the binary endpoint is considered more important and $(\omega_b = 0.75, \omega_s = 0.25)$.

If both endpoints are equally important and there is treatment effect in only one of them, the empirical powers using the $\mathcal{L}$-statistics take values between the power would have had using the two individual statistics. Given that the Bonferroni procedure assigns more importance to the more highly significant of the endpoints

(Pocock et al., 1987), the powers are in this case higher using Bonferroni than using $\mathcal{L}$-statistics.

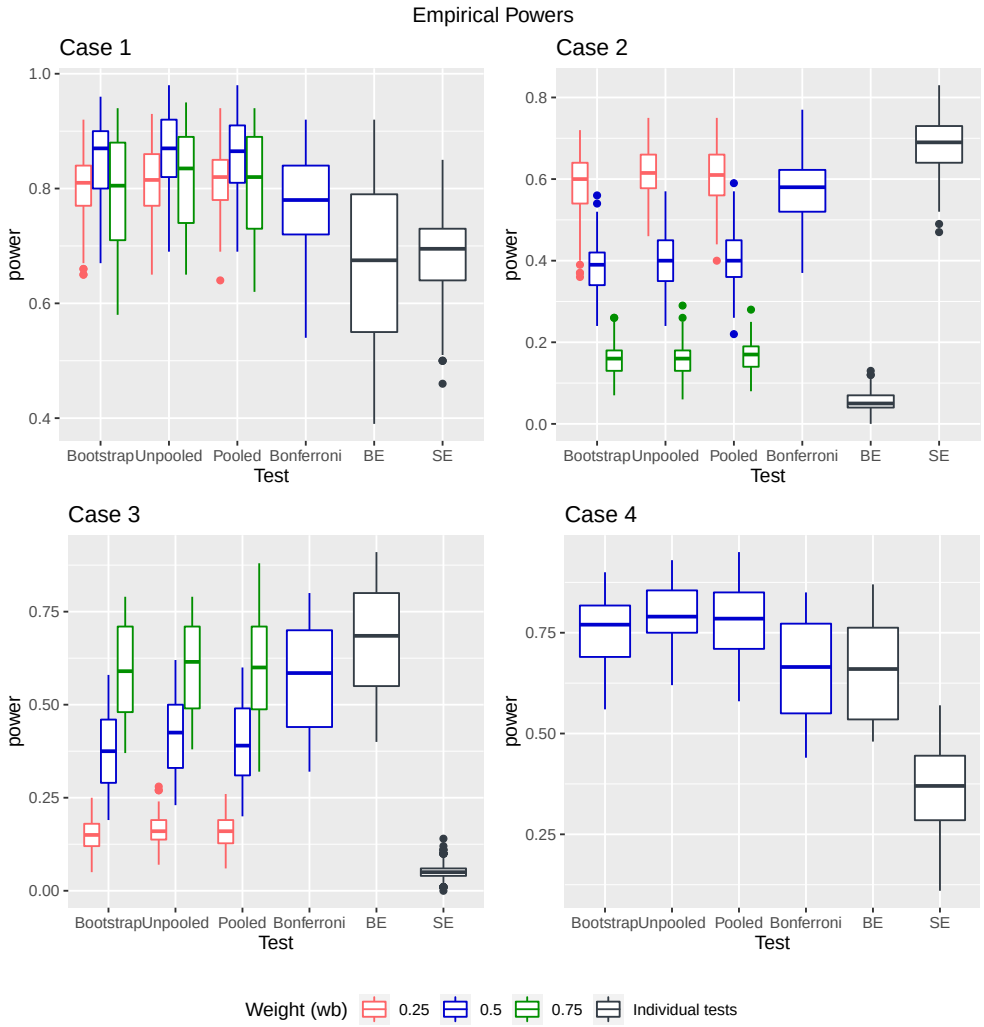


Fig. 4.3 Boxplot of empirical powers based on scenarios in Table 4.1. The empirical powers are calculated using: the $\mathcal{L}$-statistics (in (4.2)) according to the pooled, unpooled, bootstrap variance estimators; the Bonferroni procedure; and the individual statistics (4.3) and (4.4). The individual statistics for the binary and survival endpoints are labeled respectively as BE and SE. The color indicates which combination of weights (ω_b, ω_s) were used: red for $(\omega_b = 0.25, \omega_s = 0.75)$; blue for $(\omega_b = 0.5, \omega_s = 0.5)$; and green for $(\omega_b = 0.75, \omega_s = 0.25)$.

We have also evaluated the empirical powers if we have had an endpoint with a small effect instead of no effect. We observe that the difference in powers between the $\mathcal{L}$-statistics and Bonferroni procedure is smaller, and that the powers using $\mathcal{L}$-statistics could be even higher than the ones using Bonferroni (see Appendix C).

Table 4.2 Median empirical size and median empirical power from 1000 and 100 replications, respectively. The empirical size and powers are calculated using: the $\mathcal{L}$-statistics (in (4.2)) according to the pooled, unpooled, bootstrap variance estimators (labeled as Pooled, Unpooled, and Boots.); and the Bonferroni procedure (Bonf.). Under the null hypothesis there is no effect on any of the endpoints ($d = 0$,HR$= 1$). Under the alternative hypothesis there is effect on both endpoints (Case 1: $d = 0.075$, HR$= 0.75$) and the effect on the survival endpoint satisfies the proportional hazards assumptions ($t_* = 0$).

		Empirical Size				Empirical Powers (Case 1)			
		Pooled	Unpooled	Boots.	Bonf.	Pooled	Unpooled	Boots.	Bonf.
(τ_b, τ)	$(0.5, 1)$	0.054	0.054	0.047	0.050	0.82	0.84	0.81	0.79
	$(1, 1)$	0.057	0.057	0.048	0.049	0.83	0.83	0.81	0.78
θ	0.001	0.054	0.056	0.048	0.049	0.82	0.83	0.81	0.78
	2	0.055	0.055	0.047	0.049	0.82	0.83	0.82	0.78
	3	0.055	0.055	0.048	0.050	0.84	0.84	0.82	0.78
$p^{(0)}$	0.1	0.055	0.056	0.048	0.051	0.88	0.89	0.87	0.84
	0.3	0.054	0.056	0.047	0.049	0.78	0.78	0.77	0.72
a	0.5	0.055	0.055	0.050	0.049	0.85	0.85	0.85	0.81
	1	0.055	0.056	0.048	0.049	0.83	0.84	0.83	0.79
	2	0.054	0.055	0.046	0.050	0.81	0.81	0.79	0.77
(ρ, γ, η)	(0,0,1)	0.055	0.054	0.048	0.049	0.84	0.84	0.82	0.79
	(0,1,1)	0.054	0.055	0.048	0.052	0.85	0.85	0.84	0.79
	(1,0,1)	0.054	0.055	0.045	0.050	0.83	0.81	0.81	0.76
	(1,1,1)	0.055	0.056	0.048	0.048	0.84	0.84	0.83	0.78

4.7.3 Size properties

The empirical results show that the type I error is very close to the nominal level $\alpha = 0.05$ across a broad range of situations. The empirical size resulted in type I errors with a median of 0.054, 0.055 and 0.048 using the unpooled, pooled

and bootstrap variance estimators, respectively. Table 4.2 summarizes the results according to the parameters of the simulation study. The results show that the $\mathcal{L}$-statistics have the appropriate size and that are not specially influenced by the selection of weights (η, ρ, γ).

We observed that when using the unpooled and pooled estimators, the empirical size is slightly larger than 0.05, especially when $\tau_b = 1$. This can be explained mainly by the number of individuals at risk at the end of the follow-up. Having a small number of individuals difficults the smooth estimation of the probability $p_N^{(i)}(t)$ in (4.7). Therefore, we recommend the use of the bootstrap variance estimator for studies with small sample sizes with long follow-ups for both the binary and survival endpoints and where the probability of observing the binary endpoint is low.

4.8 Discussion

We have proposed a class of statistics for a two-sample comparison based on two different outcomes: one dichotomous taking care, in most occasions, of short term effects, and a second one addressed to detect long term differences in a survival endpoint. Such statistics test the equality of proportions and the equality of survival functions. The approach combines a score test for the difference in proportions and a Weighted Kaplan-Meier test-based for the difference of survival functions. The statistics are fully non-parametric and α level for testing the null hypothesis of no effect on any of these two outcomes. The statistics in the $\mathcal{L}$-class are appealing in situations when both outcomes are relevant, regardless of how the follow-up periods of each outcome are, and even when the hazards are not proportional with respect to the time-to-event outcome or in the presence of delayed treatment effects, albeit the survival curves are supposed not to cross.

We have incorporated weighted functions in the $\mathcal{L}$-statistics in order to control the relative relevance of each outcome and to specify the type of survival differences that may exist between groups. In our proposed statistics, the weights (ω_b, ω_s) have been defined with the goal of incorporating the potential difference in clinical importance between the binary and survival endpoint, and therefore they must be fixed in the planning stage. As shown in the simulation study, the power of the trial will depend on the trial objectives and then on the relevance of each of the endpoints by means of (ω_b, ω_s). The extension of these statistics

incorporating data-driven weights to maximize the power will be considered in future works.

The testing procedure using the $\mathcal{L}$-class of statistics satisfies a property called coherence that says that the nonrejection of an intersection hypothesis implies the nonrejection of any sub-hypothesis it implies, i.e., $H_{s,0}$ and $H_{b,0}$ (Romano and Wolf, 2005). However, the testing procedure based on the $\mathcal{L}$-class of statistics does not fulfil the consonant property that states that the rejection of the global null hypothesis implies the rejection of at least one of its sub-hypothesis. Bittman et al. (2009) faced the problem of how to combine tests into a multiple testing procedure for obtaining a procedure that satisfies the coherence and consonance principles. An extension of this work to obtain a testing procedure that satisfies both properties could be an important research line to consider.

This work has been restricted to those cases in which censoring does not prevent to assess the binary endpoint response. We are currently working on a more general censoring scheme where the binary endpoint could be censored. Last but not least, extensions to sequential and adaptive procedures in which the binary outcome could be tested at more than one time-point remain open for future research.

Chapter 5
Software

In this chapter we present the original software contributions that have been made throughout this thesis. The contents of this chapter have been partly published accompanying the corresponding methodologies in:

A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components.
Bofill Roig, M., and Gómez Melis, G.
Statistics in Medicine. Volume 38, Issue 11, 20 May 2019, Pages 1935–1956.
DOI: 10.1002/sim.8092.

Selection of composite binary endpoints in clinical trials.
Bofill Roig, M., and Gómez Melis, G.
Biometrical Journal. Volume 60, Issue 2, March 2018, Pages 246-261.
DOI: 10.1002/bimj.201600229.

A class of two-sample nonparametric statistics for binary and time-to-event outcomes.
Bofill Roig, M., and Gómez Melis, G.
arXiv:2002.01369 [stat.ME]

Decision tool and Sample Size Calculator for Composite Endpoints.
Bofill Roig, M., Cortés Martínez, J., and Gómez Melis, G.
arXiv:2001.03396 [stat.AP]

The chapter is twofold. First, in Section 5.1, we present CompARE, a web-based tool for designing clinical trials with composite endpoints, and its corresponding R package. Second, in Section 5.2, we present the `SurvBin` package in which we have implemented the $\mathcal{L}$-statistics presented in Chapter 4.

5.1 CompARE

We present CompARE, a comprehensive and freely available web-tool intended to provide guidance on how to deal with composite endpoints in the planning stage of a randomized controlled trial.

CompARE was initially created aiming to offer an easy implementation of the Asymptotic Relative Efficiency (ARE) method for time-to-event endpoints according to the seminal paper by Gómez and Lagakos (2013). From then on, CompARE has been continuously updated and broadened: first by adding the CompARE version for binary endpoints; later incorporating further capabilities, such as the sample size calculation. Nowadays, CompARE can be useful for different purposes as shown in Figure 5.1. In particular, it can be used to:

1. Choose the best primary endpoint to lead the trial. CompARE computes the Asymptotic Relative Efficiency method (Gómez and Lagakos, 2013; Bofill and Gómez, 2018; Bofill et al., 2020), which quantifies differences in the efficiency of using –as the primary endpoint– a composite endpoint over one of its components.
2. Specify the treatment effect for the composite endpoint based on the marginal information of the composite components and to study the performance of the composite parameters according to these. In a survival trial, it can also evaluate the proportional hazards assumption for the composite endpoint.
3. Determine the sample size for different situations, such as when the association between composite components is unknown or when the hazards are not proportional.
4. Calculate and interpret the different association measures among the composite components.

CompARE was originally build using the software Tightly Integrated Knowledge Infrastructure (Tiki Wiki CMS/groupware, 2020), but was subsequently moved to R Shiny. More recently, we have launched the R package `CompARE` which includes the R functions upon which the web-tool CompARE is built.

In this section, we first describe the R functions in the `CompARE` package. Second, we detail the basic features of the web-tool CompARE and describe how to use it and get the results. From now on, we focus on the CompARE for binary endpoints whose contents belong to this thesis. In particular, we show how the methodologies presented in Chapters 2 and 3 can be easily computed through CompARE.

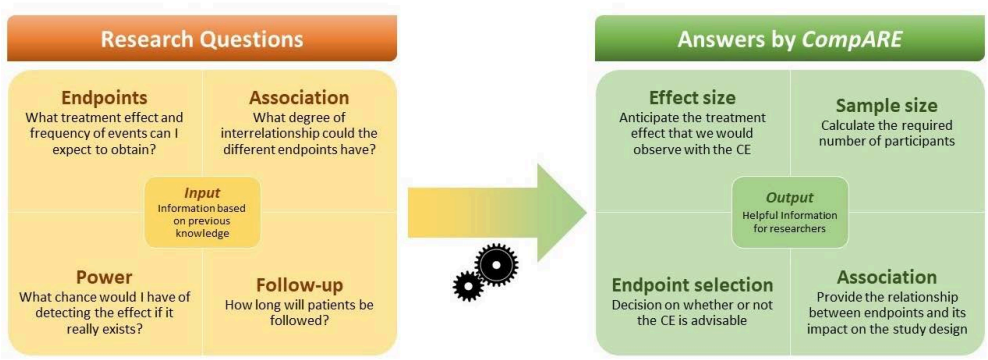


Fig. 5.1 Scheme with the inputs to be provided by the researcher and the outputs returned by the application. Link to the CompARE homepage: `http://cinna.upc.edu/compare/`. Open-source code is available at: `https://github.com/MartaBofillRoig/CompARE`.

5.1.1 `CompARE` R Package

In this section, we describe the R package upon which the web-tool CompARE has been built. The package is available on GitHub at:

`https://github.com/CompARE-Composite/CompARE-package`.

The `CompARE` package contains six functions: `prob_cbe`, `lower_corr`, `upper_corr`, `effect_cbe`, `sample_size_cbe`, and `ARE_cbe`. Table 5.1 gives an overview of these functions and relates them to the contents of this thesis and to the capabilities of the CompARE web-tool.

For the remainder of this section, we characterize each function and briefly explain the method implemented within it. As it will be shown, most functions in the `CompARE` package take common arguments. We summarize these arguments and their descriptions in Table 5.2.

Probability of the composite endpoint

This function calculates the probability of the composite of two events ε_1 and ε_2. This probability is calculated by means of the probabilities of each event and the correlation between them (Bahadur, 1961; Sozu et al., 2010; Bofill and Gómez, 2018) (see Section 3.3).

Table 5.1 R functions included in `CompARE` package along with the corresponding description and the CompARE web-tool's tab where the function is used.

Function	Description	CompARE web-tool
`prob_cbe`	Computes the probability of the composite of two events.	Summary
`lower_corr`	Computes the lower limit for Pearson's correlation.	Association Measures
`upper_corr`	Computes the upper limit for Pearson's correlation.	Association Measures
`effect_cbe`	Computes the treatment effect for composite binary endpoint (Bofill and Gómez, 2019)	Effect size
`samplesize_cbe`	Computes the sample size for composite binary endpoint (Bofill and Gómez, 2019).	Sample size
`ARE_cbe`	Computes the ARE method for composite binary endpoint (Bofill and Gómez, 2018).	Endpoint selection

Table 5.2 Arguments of the functions included in `CompARE` package and their corresponding description. Denoting by ε_1 and ε_2 two binary endpoints and by ε_* the composite endpoint defined as $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$.

Argument	Description
`p0_e1`	Probability of occurrence ε_1 in the control group (Numeric parameter)
`p0_e2`	Probability of occurrence ε_2 in the control group (Numeric parameter)
`eff_e1`	Anticipated effect for ε_1 (Numeric parameter)
`effm_e1`	Effect measure used for ε_1 (Character)
`eff_e2`	Anticipated effect for ε_2 (Numeric parameter)
`effm_e2`	Effect measure used for ε_2 (Character)
`effm_ce`	Effect measure used for the composite endpoint (Character)
`rho`	Pearson's correlation between ε_1 and ε_2 (Numeric parameter)
`alpha`	Type I error (Numeric parameter)
`beta`	Type II error (Numeric parameter)
`unpooled`	Variance estimate used for the treatment effect (Logical argument)

The function `prob_cbe` can used by means of:

```
prob_cbe(p_e1,p_e2, rho)
```

where `p_e1` and `p_e2` denote the probabilities of ε_1 and ε_2, respectively; and `rho` is the correlation between ε_1 and ε_2.

Correlation Bounds

Pearson's correlation between two binary outcomes takes values between two bounds defined according to the probabilities of the binary outcomes. The functions `lower_corr` and `upper_corr` calculate the lower and upper bounds of Pearson's correlation based on the probabilities of two binary outcomes (Bahadur, 1961; Bofill and Gómez, 2018) (see Section 3.3).

These functions can be called by:

```
lower_corr(p_e1,p_e2)
```

and

```
upper_corr(p_e1,p_e2, rho)
```

where `p_e1` and `p_e2` denote the probabilities of two binary endpoints.

Effect size for composite endpoints

We have implemented the calculation of the effect size for composite endpoints in the function `effect_cbe`. The composite endpoint is assumed to be the combination of two events (ε_1 and ε_2). We compute the effect size on the basis of anticipated information of the composite components and the correlation between them as it is explained in Chapter 2 (see Sections 2.2 and 2.6).

The function `effect_cbe` can be called by:

```
effect_cbe(p0_e1, p0_e2,
           eff_e1, effm_e1, eff_e2, effm_e2,
           effm_ce ="diff", rho)
```

where `p0_e1` and `p0_e2` denote the probabilities of ε_1 and ε_2 in the control group, respectively; `eff_e1` and `eff_e2` are the anticipated effects for the events ε_1 and ε_2, respectively; `rho` is Pearson's correlation between ε_1 and ε_2.

The effects for ε_1 and ε_2 can be anticipated in `eff_e1` and `eff_e2` by means of the difference of proportions, risk ratio, and odds ratio. The arguments `effm_e1` and `effm_e2` can be used for specifying the effect measure preferred. We will

use `effm_e1` = "diff" for difference of proportions, `effm_e1` = "rr" for risk ratio, `effm_e1` = "or" for odds ratio (similarly for `effm_e2`). Also, using the argument `effm_ce`, we specify the effect measure we are interested in for the composite endpoint.

Sample size for composite endpoints

We have implemented the sample size calculation for composite binary endpoints in the function `samplesize_cbe`. The primary endpoint is assumed to be a composite binary endpoint formed by combining two events (ε_1 and ε_2). The sample size is computed for evaluating differences between two groups in terms of the risk difference, risk ratio or odds ratio. We calculate the sample size based on anticipated information of the composite components and the correlation between them as it is explained in Chapter 2 (see Sections 2.4 and 2.6).

The function `samplesize_cbe` can be called by:

```
samplesize_cbe(p0_e1, p0_e2,
               eff_e1, effm_e1, eff_e2, effm_e2,
               effm_ce ="diff",
               rho, alpha = 0.05, beta = 0.2,
               unpooled = TRUE)
```

where `p0_e1` and `p0_e2` denote the probabilities of ε_1 and ε_2 in the control group, respectively; `eff_e1` and `eff_e2` are the anticipated effects for ε_1 and ε_2, respectively; `rho` is Pearson's correlation between ε_1 and ε_2; `alpha` and `beta` are the type I and type II errors, respectively; and `unpooled` denotes the variance estimate used for the sample size calculation ("TRUE" for unpooled variance estimate, and "FALSE" for pooled variance estimate).

The effects for ε_1 and ε_2 can be anticipated in terms of the difference of proportions, risk ratio, and odds ratio as before by means of the arguments `effm_e1` and `effm_e2`. Using the argument `effm_ce`, we specify the effect measure for the composite endpoint.

Endpoint selection for composite endpoints

We have implemented the Asymptotic Relative Efficiency (ARE) method for binary composite endpoints in the function `ARE_cbe`. The composite endpoint is assumed to be the combination of two events (ε_1 and ε_2), and additionally we assume that there is one endpoint that is more relevant than the other. We consider

ε_1 the most relevant endpoint and ε_2 the additional one. The ARE gives then a criterion to decide whether to use a composite binary endpoint (ε_*) or to use its more relevant component (ε_1) as the primary endpoint to lead the study. We compute the ARE method in terms of the odds ratio as well as in terms of the risk difference according to the methodology explained in Chapter 3.

The function `ARE_cbe` can be called by:

```
ARE_cbe(p0_e1, p0_e2,
        eff_e1, effm_e1, eff_e2, effm_e2,
        effm_ce = "or", rho)
```

where `p0_e1` and `p0_e2` denote the probabilities of ε_1 and ε_2 in the control group, respectively; `eff_e1` and `eff_e2` are the anticipated effects for ε_1 and ε_2, respectively; and `rho` is Pearson's correlation between ε_1 and ε_2.

The effects for ε_1 and ε_2 can be anticipated in terms of the difference of proportions, risk ratio, and odds ratio as before using the arguments `effm_e1` and `effm_e2`.

5.1.2 Web-tool CompARE

The web-tool CompARE is a completely free web platform that can be used as a tool for clinicians, medical researchers and statisticians. All users can access through a standard web browser using the web address `https://cinna.upc.edu/compare/`.

CompARE is built for clinical trials with multiple endpoints of interest and, in particular, with composite endpoints. Specifically, CompARE is appropriate for trials that meet the following conditions:

- **Two-arm design**. Studies aimed at comparing two different groups.
- **Superiority design**. Studies designed to establish whether a new intervention is superior to the standard care.
- **Endpoints under study**. Studies with the following endpoints of interest: two binary endpoints, ε_1 and ε_2, and the composite endpoint defined as the event that occurs whenever one of the endpoints ε_1 and ε_2 is observed, that is, $\varepsilon_* = \varepsilon_1 \cup \varepsilon_2$.

The basic structure of the user interface of CompARE is schematized in Figure 5.2. The user interface of CompARE is composed of three parts: Input panel, Menu bar, and Output panel. The workflow within CompARE relates these three parts with each other as follows: the first step is to enter the input parameters for the

computations in the left panel (Input panel) according to available information based on previous knowledge; the second step is to select the output the user is interested in by using the menu at the top (Menu bar); the third and final step is to examine the results and use the explanation boxes to understand them (in Output panel). Next, we explain in detail these three parts that compose the webtool CompARE.

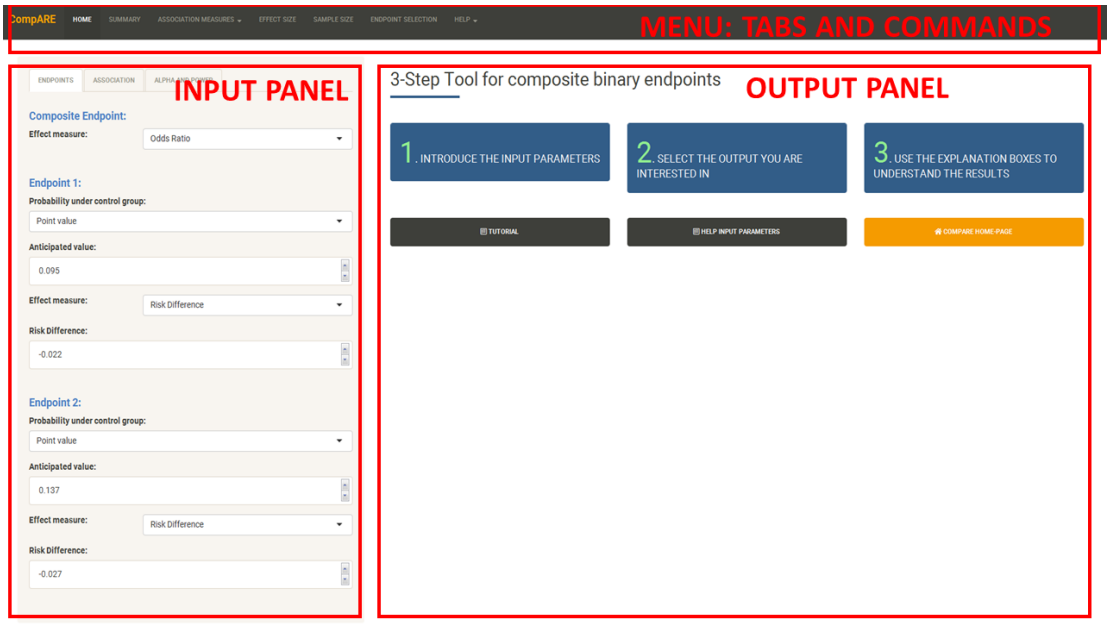


Fig. 5.2 Structure of CompARE: Input panel where users introduce the values; Menu bar where capabilities of CompARE are exposed; and Output panel where results and plots are showed.

Input panel

The user is prompted for the anticipation of different parameter values in the Input panel. The Input panel of CompARE-Binary is composed of three tabs: (i) Endpoints; (ii) Association; and (iii) Alpha and beta. We exemplify the Input panel in Figure 5.3.

(i) *Endpoints*: This is the main part of the Input panel in which users should anticipate information about the composite endpoint ε_* and the composite

components ε_1 and ε_2. We list below the parameters that the users should provide.

Required information about the composite endpoint:
- Effect measure: Measure for quantifying the effect for the composite endpoint. The user can choose from the options: Odds ratio, Risk difference, and Risk ratio.

Required information about each composite component:
- Effect measure: Measure for quantifying the effect for the endpoint 1 (analogously for endpoint 2).
- Probability under control group: whether the user anticipates a point value or an interval of plausible values for the probability of observing the event for the endpoint 1 in the control group (analogously for endpoint 2).
  - If "Point value" has been chosen, then the user should provide the anticipated value for the probability of observing the event for the endpoint 1 in the control group.
  - If "Interval plausible values" has been chosen, then the user should provide the lower and upper values between which the probability of the endpoint 1 in the control group takes values on.
- Effect measure: Measure for quantifying the effect for the endpoint 1 (analogously for endpoint 2). The user can choose between: Odds ratio, Risk difference, and Risk ratio.
  - Effect value: Expected effect size on the Endpoint 1 according to the effect measure previously chosen (Odds ratio, Risk difference, or Risk ratio).

(ii) *Correlation*: In the second tab of the Input panel, the users should set the strength of correlation between endpoints by means of Pearson's correlation coefficient.

Remarks:
- The correlation is bounded and its bounds depend on the marginal parameters. See Association Tab (in the Menu bar) for more information.
- In CompARE, the correlation is assumed to be equal between groups.
- Since in practice most of the times this information is unavailable, CompARE will produce plots to visualize how much the correlation impacts on the calculations.

(iii) *Alpha and Power*: In third tab of the Input panel, the user can enter some additional information for the sample size calculation.

Required information:
- Significance level: Type I error considered for the study.
- Power: Statistical power considered for the study.
- Variance of the effect: Formula used for estimating the variance of the effect. The user can choose between the unpooled variance estimator and the pooled variance estimator.

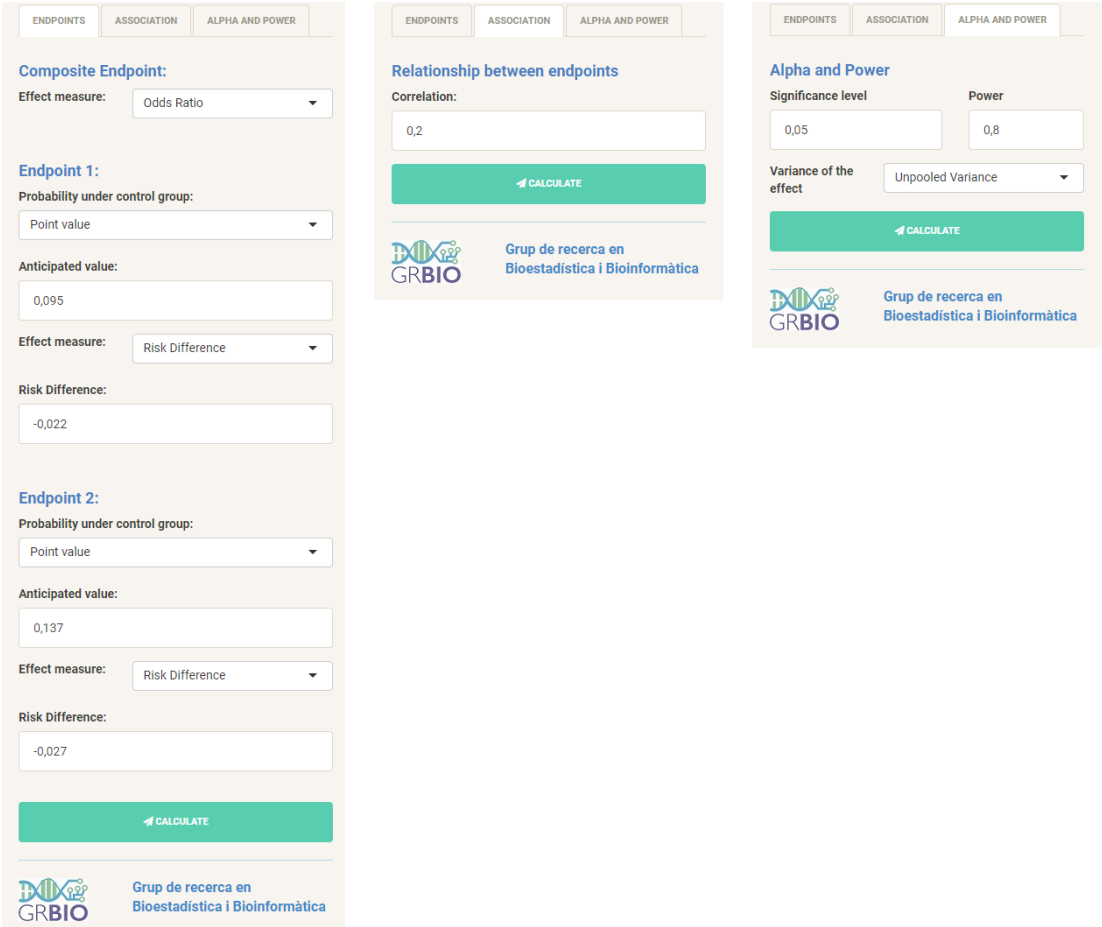


Fig. 5.3 Input panel of CompARE.

Menu bar and Output panel

Throughout the Menu bar, the users can select the calculations they are interested in. Immediately after that, the results are shown in the Output panel. An example of the results displayed in the Output panel can be seen in Figure 5.4. In this section, we explain all available options in the Menu bar together with the outputs that this provides.

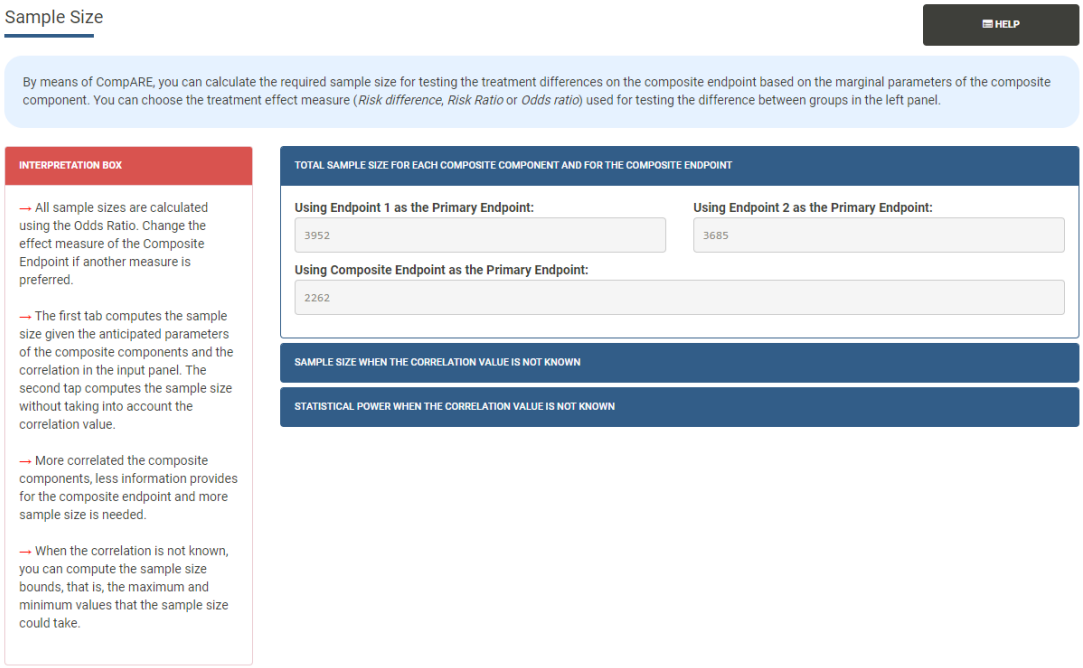


Fig. 5.4 Example of Output panel.

- *Summary*: In the first tab, CompARE provides a key messages about the trial design as well as a short report according to the parameters anticipated in the Input panel.

  Outputs obtained:
  · Answer for several key questions, such as: "How much sample size is needed?" or "What is the expected odds ratio for the composite endpoint?".
  · Summary for the composite endpoint.

  · Summary for the composite components.

- *Association measures*: When using a composite binary endpoint, one needs to take into account the degree of association between the composite components. However, this association is usually unknown or difficult to anticipate.
  In CompARE, you can calculate which values the association between the components could have based on the marginal parameters. The association can be specified using Pearson's correlation measure or Relative Overlap.

  Outputs obtained:
  · Lower and upper bounds for the correlation.
  · Probability of the composite endpoint.
  · Probability of the overlap: probability of the intersection of both outcomes.
  · Relative overlap: the ratio between the probability of the intersection and the probability of the composite endpoint.

- *Effect Size*: Before a study is conducted, investigators need to anticipate which is the minimum effect relevant to be detected. Depending on this effect, analysis and sample size will be carried out.
  By means of CompARE, you can calculate the effect size for a composite binary endpoint based on the marginal parameters of its components.

  Outputs obtained:
  · Effect Size for each composite component.
  · Effect Size for the composite endpoint given the correlation value.
  · Effect Size for the composite endpoint given the correlation category.

- *Sample Size*: When designing a study, investigators need to determine how many subjects should be included. By enrolling too few subjects, a study may not have enough statistical power to detect a treatment difference. Enrolling too many patients can be unnecessarily costly or time-consuming.
  To size a trial with a composite binary endpoint, one needs to specify the event rates and the effect sizes of the composite components along with the correlation between them. In practice, the marginal parameters of the components can be obtained from previous studies or pilot trials, however, the correlation is often not previously reported and thus usually unknown.
  By means of CompARE, the user can calculate the required sample size for composite binary endpoints based on the anticipated information on the components in cases where the correlation is totally or partially known, as well as where there is uncertainty in the event rate values of the components.

Outputs obtained:
· Sample Size for each composite component.
· Sample Size for the composite endpoint given the correlation value.
· Sample Size for the composite endpoint given the correlation category.

- *Endpoint Selection*: CompARE can help you to make a more informed decision on the Primary Endpoint you should use in your study.
  Considering the Endpoint 1 as the most relevant endpoint you must use in your trial to test the treatment efficacy, CompARE will help you to choose between Endpoint 1 (ε_1) and the Composite Endpoint (ε_*) as the primary endpoint for the trial.

  Outputs obtained:
  · ARE criteria for deciding between ε_1 and ε_*.
  · ARE criteria according to the correlation value.

- *Help*: CompARE includes a tutorial to guide users through the software and whereby details of the methods on which CompARE is based are provided.

5.2 `SurvBin` R Package

We developed the R package `SurvBin` to facilitate the computation of the $\mathcal{L}$-statistics presented in Chapter 4, and made it available on GitHub at:

`https://github.com/MartaBofillRoig/SurvBin`.

The `SurvBin` package offers several tests for the comparison of two populations: the $\mathcal{L}$-statistic for comparisons based on binary and time-to-event outcomes; the score statistic for binary outcomes; and Kaplan-Meier based-tests for time-to-event outcomes. The `SurvBin` package also provides additional functions for computing the covariance between binary and survival statistics, and for simulating bivariate binary and survival data. Table 5.3 gives an overview of these functions and their capabilities. Table 5.4 summarizes the arguments used in the functions in the `SurvBin` package.

Table 5.3 R functions included in `SurvBin` package along with the corresponding description and the methods implemented. Theorems are available in Appendix C.

Function	Description	Methods
`lstats`	Compute standardized $\mathcal{L}$-statistics	Section 4.2 (Theorem 4.4)
`lstats_boots`	Compute standardized $\mathcal{L}$-statistics using the bootstrap variance estimator	Sections 4.5 and 4.7
`bintest`	Compute univariate binary statistics	Section 4.2 (Theorem 4.3)
`survtest`	Compute univariate survival statistics	Section 4.2 (Theorem 4.3)
`survbinCov`	Compute the covariance between binary and time-to-event statistics	Section 4.3 (Theorem 4.3)
`simsurvbin`	Simulate bivariate binary and survival data	Section 4.7

Table 5.4 Arguments of the functions included in `SurvBin` package and their corresponding description.

Argument	Description
`time`	Vector of the right-censored data
`status`	Vector of the status indicator
`binary`	Vector of the binary data
`treat`	Vectors of the treatment group indicator
`tau0, tau, taub`	Follow-up configuration
`wb, ws`	Scalar parameters that controls the weight ω
`rho, gam, eta`	Scalar parameters that controls the weight $\hat{Q}(t)$
`var_est`	Variance estimate to use (`pooled` or `unpooled`).

5.2.1 R functions in SurvBin package

Herein, we describe the functions of the SurvBin package and their usage.

$\mathcal{L}$-statistics

The function lstats computes the $\mathcal{L}$-statistics presented in Chapter 4. The call to the function lstats is as follows:

```
lstats(time, status, binary, treat,
       tau0, tau, taub,
       rho, gam, eta, wb, ws,
       var_est)
```

where time, status, binary and treat are vectors of the right-censored data, the status indicator, the binary data and the treatment group indicator, respectively; tau0, tau, taub denote the follow-up configuration; wb, ws are the weights ω; rho, gam, eta are scalar parameters that controls the weight $\hat{Q}(t)$ which is given by $\hat{Q}(t) = \hat{G}(t-)^{\eta} \cdot \hat{S}(t-)^{\rho} \cdot (1 - \hat{S}(t-))^{\gamma}$; and var_est indicates the variance estimate to use (pooled or unpooled).

As a result, lstats returns a list consisting of:

(i) the standardized $\mathcal{L}$-statistic given in (C.4), $\mathbf{U}_n^{\omega}(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$;
(ii) the $\mathcal{L}$-statistic given in (4.2), $\mathbf{U}_n^{\omega}(\hat{Q})$;
(iii) the standard deviation of the $\mathcal{L}$-statistic given in (4.9).

We have also included in the SurvBin package the function lstats_boots. This function computes the standardized $\mathcal{L}$-statistics by using the bootstrap variance estimator. The function lstats_boots can be used by means of:

```
lstats_boots(time, status, binary, treat,
             tau, rho, gam, eta,
             wb, ws, Boot)
```

where time, status, binary,treat, and tau, rho, gam, eta, wb, ws are the same parameters that we have in lstats; and where Boot denotes the number of bootstrap samples.

Binary statistics

The function `bintest` performs a test statistic for testing the difference in two population proportions. The call to the function `bintest` is as follows:

```
bintest(binary, treat, var_est)
```

where `binary` and `treat` are vectors of the binary data and the treatment group indicator, respectively; and `var_est` indicates the variance estimate to use (`pooled` or `unpooled`).

The output provided by `bintest` returns a list consisting of:

(i) the standardized binary statistic $U_{b,n}(\tau_b)/\hat{\sigma}_b$;
(ii) the binary statistic given in (4.3), $U_{b,n}(\tau_b)$;
(iii) the standard deviation of the binary statistic given in (4.10).

Survival statistics

The function `survtest` performs a test for right-censored data. It uses the Weighted Kaplan-Meier family of statistics (see (4.4)) for testing the differences of two survival curves.

```
survtest(time, status, treat, tau, rho, gam, eta, var_est)
```

where `time, status` and `treat` are vectors of the right-censored data, the status indicator and the treatment group indicator, respectively; `tau` denote the end of follow-up; `wb, ws` are the weights ω; `rho, gam, eta` are scalar parameters that controls the weight $\hat{Q}(t)$ which is given by $\hat{Q}(t) = \hat{G}(t-)^{\eta} \cdot \hat{S}(t-)^{\rho} \cdot (1 - \hat{S}(t-))^{\gamma}$; and `var_est` indicates the variance estimate to use (`pooled` or `unpooled`).

As a result, `survtest` returns a list consisting of:

(i) the standardized survival statistic, $U_{s,n}(\tau_0, \tau; \hat{Q})/\hat{\sigma}_s$;
(ii) the survival statistic given in (4.4), $U_{s,n}(\tau_0, \tau; \hat{Q})$;
(iii) the standard deviation of the survival statistic given in (4.11).

Covariance computation

The function `survbinCov` calculates the estimator of the covariance between the binary and survival statistics, $U_{b,n}$ and $U_{s,n}(Q)$, defined by (4.3) and (4.4). It computes the covariance estimator $\hat{\sigma}_{bs}$ given in Section 4.3.3.

The function `survbinCov` can be called by:

```
survbinCov(time, status, binary, treat,
           tau0, tau, taub,
           rho, gam, eta,
           var_est)
```

where `time, status, binary` and `treat` are vectors of the right-censored data, the status indicator, the binary data and the treatment group indicator, respectively; `tau` denote the end of follow-up; `wb, ws` are the weights ω; `rho, gam, eta` are scalar parameters that controls the weight $\hat{Q}(t)$ which is given by $\hat{Q}(t) = \hat{G}(t-)^{\eta} \cdot \hat{S}(t-)^{\rho} \cdot (1-\hat{S}(t-))^{\gamma}$; and `var_est` indicates the variance estimate to use (`pooled` or `unpooled`).

In this work, we estimate $\lambda^{(i)}_{X,T}(t)$ (see definition in Theorem 4.3 in Appendix C) by means of the Epanechnikov kernel function, and the local bandwidth selection and the boundary correction described by Muller and Wang (1994) by using the `muhaz` package (Hess and Gentleman, 2019).

Simulating bivariate binary and time-to-event data

The function `simsurvbin` simulates bivariate binary outcomes and survival times. The simulation is based on a copula-based framework and uses the conditional sampling described in (see Appendix A of Trivedi and Zimmer, 2005 for further information). The function `simsurvbin` can be called by:

```
simsurvbin(a.shape, b.scale, rate.param,
           prob0, ass.par, ss,
           censoring="Exp")
```

where `a.shape` and `b.scale` are the shape and scale parameters for the Weibull distribution; `rate.param` is the distributional parameter for the censoring; `prob0` probability binary outcome; `ass.par` denotes the association between the binary and time-to-event outcomes according to a Frank's copula; and `ss` is the sample size per arm.

Throughout the argument `censoring`, the user can choose between uniform and exponential distributions for the censoring distribution. Depending on that, the parameter `rate.param` will correspond to the rate for the exponential distribution (say c where $\text{Exp}(c)$) or the maximum for the uniform (say c where $\text{Unif}(0, c)$).

Chapter 6
Conclusions and Future Research

This thesis addresses the design and analysis of trials with multiple endpoints. The thesis is mainly divided into two topics. In the first, we coped with the design of trials with composite binary endpoints, going from the effect size and sample size calculation to the selection of the components for a primary composite endpoint. In the second, we proposed new methods for comparing two groups in seamless phase II/III trials using binary and survival endpoints. We look forward to continuing the work started in this thesis and adapting it to current problems. In this chapter, we summarize the main conclusions of this thesis and outline some future lines of research.

6.1 Composite endpoints

Despite being widely used, composite endpoints entail challenges in both designing trials and interpreting results. In Chapter 2, we have shown that calculating the sample size for composite binary endpoints needs more than the anticipated effect size and event rates of the composite components; it also needs the correlation between them. We demonstrated that the sample size increases as the correlation does and that it strongly depends on the correlation value. We proposed strategies for deriving the sample size when the correlation is not specified. The strategies are based on the stratification of the correlation into different categories. We did so by splitting the rank of the correlation into three equal-sized intervals, and then considering three correlations categories: weak for the interval whose correlation values are lower; moderate for those intermediate correlation values; and strong for those correlation values that are higher. The sample size is then calculated using the maximum correlation value across the correlation category. We showed

in the simulation study in Section 2.7 that the sample size strategies assure the pre-specified power even without previous knowledge on the correlation.

In Chapter 3, we have proposed the Asymptotic Relative Efficiency (ARE) method to quantify the gain in efficiency of using a composite binary endpoint instead of its most relevant component as primary endpoint to lead the trial. The ARE method allows to make an informed decision by providing a criteria to choose between these two endpoints. The method relies on the magnitude of the treatment effects and the event rates of the composite components, and the correlation between them. We discussed the influence that these parameters have on the decision throughout a numerical study and summarized the conclusions into statistical guidelines. Although composite endpoints are widely used as primary endpoints in clinical trials, we noticed that they are not always the best option.

In summary, we have shown that the design of a randomized controlled trial involving a composite endpoint needs a careful specification of the expected rates, the treatment effects and the correlation. While the parameters of the composite components can often be derived from previous studies, the correlation are seldom disclosed. In this thesis, we have seen that the correlation between the composite components plays an important role in both selecting the primary endpoint –by means of the ARE method– and calculating the appropriate sample size. Therefore, it is of utmost importance to consider how much the components of the composite endpoint are associated, to quantify this association and to report it in all documents derived from the trials.

The methodologies presented in Chapters 3 and 4 are focused solely on studies with composite endpoints with two components, but they might be extended as follows. The sample size formula for composite endpoints with more than two components can be straightforwardly obtained following the same rationale that in Bofill and Gómez (2019). However, the difficulty in this case will rely on the fact that the underlying correlation structure will become increasingly complex with the rising number of components. This complicates the study of the behavior of the sample size in terms of the correlation, and makes the assumption of equal correlation structure between groups even stronger than it already was.

In studies with more than two composite components and where there exists an order of importance of such components, the ARE method can be recursively used to decide whether to add more endpoints into the composite primary endpoint. Extensions of the method that could select the most efficient components for the composite endpoint are left for future research.

Furthermore, extensions to unbalanced designs can be considered as well in future works. Despite the fact that sample size formulae for composite endpoints

under unbalanced designs can be easily obtained from Bofill and Gómez (2019), it would be interesting to assess the impact of the unbalanced settings in keeping the type I error, especially in studies with small sample sizes.

The ARE method has been done accounting for unbalanced designs. However, the method assumes that the designs for both relevant and composite endpoints have the same proportion of patients allocated in the control group. It might be interesting to evaluate whether having different allocation proportions in the relevant and composite endpoints affects the decision of the primary endpoint.

We have presented in Chapter 5 the implementations of the previous mentioned methodologies in the web-tool CompARE and its corresponding R package. In the forthcoming future, we plan to continue working on CompARE, upgrading it in several directions.

A good data visualization may give at first glance more information to users than some summary statistics and may aid the newcomer to understand what CompARE is offering quickly. We plan to improve the plots in CompARE by using `ggplot` (Wickham and Grolemund, 2020) and make them interactive by using `plotly` (Sievert, 2020). So that the plots can be for instance zoomed and exported by the user. Also, some functions for plotting R objects will be added into the R package `CompARE`.

Another feature we would like to add in CompARE is the option of generating dynamic reports. The idea here is to allow users to download summary reports after using CompARE. We are going to explore the possibility of incorporating add-ons for the generation of reports by means of `rmarkdown` (Wickham, 2020b; Xie et al., 2020).

Last but not least, we plan to enhance and complete the R package `CompARE` and make it available from the CRAN (`http://cran.r-project.org`).

6.2 Binary and survival endpoints

Many biomedical studies are conducted to compare a treatment group with a control group on the basis of several hypotheses. In Chapter 4, we have proposed a class of statistics for comparing two groups based on binary and time-to-event outcomes. The statistics are appealing in a broad life-studies situations and, specially, in cancer immunotherapy trials where both binary and time-to-event endpoints are of interest and where the proportional hazards assumption is rarely met.

We based the proposed class of statistics on a weighted combination of a difference in proportions test and a weighted Kaplan-Meier test-based for the difference of survival functions. The statistics are fully non-parametric and do not need the proportional hazards assumption for the survival outcome. Moreover, our proposal adds versatility into the study by incorporating random weights and flexibility by allowing for different follow-up configurations for the binary and survival outcomes. We presented our proposal together with the R package `SurvBin` where the methodology has been implemented.

As mentioned in the discussion in Section 4.8, the major limitation of this work is that the methodology does not accommodate the situation where censoring may prevent the observation of the binary outcome. We are aware of the importance of extending the methodology to encompass studies with complex censoring schemes between the binary and time-to-event outcomes, and we plan to work on this issue in the near future. A second limitation is that there may be a semi-competing risk problem between the binary and time-to-event outcomes. In clinical practice, it is common to assign the "worse" response to those patients that died before the evaluation of the binary response. For example, if a patient died before evaluating tumor progression, it is usually considered that the patient had progression, albeit it was not observed. Further work should be done to evaluate the implications of this practice to guide better use.

As future research, we would like to study potential extensions for sequential monitoring designs. Besides, alternatives approaches for estimating the covariance between the binary and survival statistics will be as well considered and incorporated into the R package `SurvBin`. We will also improve the `SurvBin` package by adding long-form guides (vignettes) and examples.

We believe that this work might be exploited and extended in several directions. Next, we go through open questions and related research lines that motivate our future work.

6.2.1 Survival by tumor response

In neoadjuvant cancer trials, the binary endpoint pathologic complete response (pCR) has been commonly used as the primary endpoint for phase II trials or even as an endpoint for accelerated approval in high-risk populations (FDA Guidance, 2014). If granted, a two-arm confirmatory trial is often required to demonstrate the efficacy with long-term endpoints, such as overall survival. However, the design of such a trial based on prior information on the pCR effect is not straightforward.

Patients whose tumors show pathological complete response (pCR) at surgery after neoadjuvant chemotherapy seem to have better long-term survival outcome than patients whose tumors do not, regardless of the treatment group they are assigned to (Cortazar et al., 2014).

Aiming at designing a phase III trial with overall survival for comparing two treatment groups using the results from previous trials with pCR outcome, we are currently working on the anticipation of the expected effect size and on the calculation of the sample size under different scenarios and based on the pCR response and the survival by response by each arm. For this purpose, we consider the distribution of overall survival as a mixture of pCR responders and non-responders in each treatment arm. We base the comparison between groups on the difference of restricted mean survival times, which gives a clinically meaningful summary of treatment effect and moreover does not rely on the proportionality of the hazards. The preprint version of this joint work with Prof. Y. Shen and Prof. G. Gómez Melis is available in Arxiv (arXiv:2008.12887 [stat.ME]).

A somehow similar situation is found in the treatment of acute leukemia. When a patient is diagnosed with acute leukemia, an induction therapy is selected aimed at reducing the total body leukemic cell population. A bone marrow biopsy is afterwards performed to evaluate the patient's response to that initial therapy. The patient is then deemed a responder, if he/she achieved the desired response, or non-responder, if not. Based on that, the clinician should typically choose the next of treatment to follow, seeking to maximize the expected benefit to the patient with respect to a time-to-event outcome, such as overall survival.

This problem has triggered the interest in sequential multiple assignment randomized trials, where patients are randomized repeatedly at each time in which the therapy could be changed or modified (Tsiatis et al., 2020). In this context, we would like to explore the extension of the statistics proposed in Chapter 4 for comparing response rates and survival after reassignment by tumor response.

Additional related topics such as the study of the association between pCR and overall survival (Broglio et al., 2016; De Michele et al., 2015), the estimation of the time-to-tumor under different scenarios (Gómez Melis, 1986; Vardi et al., 2001; Weedon-Fekjaer et al., 2008) and its evaluation as surrogate of overall survival (Burzykowski and Buyse, 2006; Buyse et al., 2018) are several directions we may explore in the future.

6.2.2 Estimands in clinical trials

Randomized controlled trials are expected to be free from confounding variables, however, in practice, certain events may occur that complicate the measurement of the endpoints associated with the clinical question and affect the interpretation of treatment effects (ICH E9(R1), 2020). These events, named as intercurrent events, include treatment discontinuation and treatment switching, among many others (Akacha et al., 2017).

The occurrence of intercurrent events may be especially worrisome in survival trials. Since intercurrent events occur between randomization and before the observation of the time-to-event endpoint, they may affect how we evaluate the scientific question. Just as an example, in the case of treatment switching, patients switch from their assigned treatment onto an alternative. This provokes that the observed time for a patient that switched will not be the one specified in the design. Would this observed time have been the same if treatment switching had not taken place? Under circumstances where the answer is yes, the trial integrity could be compromised (Rufibach, 2018).

The ICH E9(R1) (2020) addendum takes the first step in drawing designs according to trial objectives and accounting for plausible intercurrent events. By means of the definition of the so-called estimands ("what to be estimated"), the addendum promotes trial designs that distinguish between: what to be estimated, how this will be estimated, and how much robust the conclusions would be according to the assumptions made in the design.

The estimands framework and the potential confounding when intercurrent events are present open the door to start using new methodologies, such as causal inference (Robins, 1986; Tsiatis et al., 2020; Hernán and Robins, 2020), in clinical trials. As future work, we will explore the estimation of the effect measured through time-to-event endpoints where intercurrent events may happen and how the design of such trials would be.

6.3 Extending Pitman's Asymptotic Relative Efficiency

In recent years, there has been an increasing interest in clinical trials designed to evaluate multiple drugs and/or multiple disease populations in parallel. Specially in cancer studies, where multiple drugs might be tested over different cancer types. The so-called master protocol describes the design of such trials (FDA Guidance,

2018). In contrast to traditional clinical trial designs, which are often limited in the questions they deal with, master protocols cover multiple substudies, which may have different objectives, and thus address different questions. However, the complex structure of master protocols leads to regulatory and statistical challenges.

Master protocols use a single infrastructure, trial design, and protocol allowing for efficient and accelerated drug development. Such trials are often classified into: (i) trials testing a single drug across multiple cancer populations, referred to as basket trial; (ii) trials testing multiple drugs in a single disease population, referred to as umbrella trials; and (iii) trials which may take the form of basket or umbrella trials and in which new substudies can potentially be added or stopped dynamically during the course of the trial, named platform trials.

In the context of studies where the full population is divided into two subpopulations in which treatment could potentially have different modes of action and different benefits, we want to explore the extension of the definition of the Asymptotic Relative Efficiency (ARE) given by Pitman (Lehmann and Romano, 2005) aiming at designing more efficient basket trials by maximizing the information we obtain for each subpopulation.

Given two statistics for the same hypothesis test, the ARE compares the asymptotic efficiencies of the statistics and can be interpreted as the ratio of the Fisher information associated to each statistic. In previous works, Gómez and Lagakos (2013) and Bofill and Gómez (2018) used the ARE to compare two non-equivalent set of hypotheses in the context of composite endpoints. Following the same rationale, we would like to consider the extension of the ARE as a key measure to quantify the information contained in a set of hypotheses. Since the concept of information is closely related to the sample size, we plan to exploit this fact in designing trials, specifically, in determining the sample size when dealing with multiple correlated hypotheses.

Appendices

Appendix A
Appendix of Bofill and Gómez (2019)

Let X_{ijk} denote the response of the k-th binary endpoint for the j-th patient in the i-th group of treatment ($i = 0, 1,\ j = 1, ..., n,\ k = 1, 2$). We denote by X_{ij*} the composite response defined as

$$X_{ij*} = \begin{cases} 1, \text{ if } X_{ij1} + X_{ij2} \geq 1 \\ 0, \text{ if else } X_{ij1} + X_{ij2} = 0 \end{cases} \tag{A.1}$$

W denote by $p_1^{(i)} = \mathrm{P}(X_{ij1} = 1) = 1 - q_1^{(i)}$, $p_2^{(i)} = \mathrm{P}(X_{ij2} = 1) = 1 - q_2^{(i)}$ and $p_*^{(i)} = \mathrm{P}(X_{ij*} = 1) = 1 - q_*^{(i)}$ the probabilities of observing each endpoint in the i-th group. Let $O_k^{(0)}, \delta_k, \mathrm{R}_k, \mathrm{OR}_k$ be the odds under the control group, the risk difference, risk ratio and odds ratio, respectively, for the k-th endpoint, that is, $O_k^{(0)} = \frac{p_k^{(0)}}{q_k^{(0)}}$, $\delta_k = p_k^{(1)} - p_k^{(0)}$, $\mathrm{R}_k = \frac{p_k^{(1)}}{p_k^{(0)}}$, and $\mathrm{OR}_k = \frac{p_k^{(1)}/q_k^{(1)}}{p_k^{(0)}/q_k^{(0)}}$. We denote by $\theta = (p_1^{(0)}, p_2^{(0)})$ the vector of marginal event rates, and $\lambda = (\delta_1, \delta_2)$ the vector of effect sizes.

Let $\rho^{(i)}$ represent the correlation between X_{ij1} and X_{ij2} in the i-th treatment group, and ρ refer to the correlation when it is assumed to be equal in both groups, i.e., $\rho = \rho^{(0)} = \rho^{(1)}$.

A.1 Derivation of the composite effect from the margins

We derive the expression for the composite treatment effect in terms of the marginal component and the correlation described in Sections 2.2 and 2.6, and we prove the monotone performance of the risk difference with respect to the correlation ρ.

Theorem A.1 (Composite effect from margins). *The composite effect for the composite endpoint can be expressed in terms of the component parameters as follows:*

(i) The risk difference for the composite endpoint, δ_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, δ_1, δ_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\begin{aligned}\delta_* = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1\delta_2 + \rho^{(0)}\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}} \\ -\rho^{(1)}\sqrt{(p_1^{(0)}+\delta_1)(p_2^{(0)}+\delta_2)(q_1^{(0)}-\delta_1)(q_2^{(0)}-\delta_2)}\end{aligned} \tag{A.2}$$

(ii) The risk ratio for the composite endpoint, R_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, R_1, R_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\mathrm{R}_* = \frac{p_1^{(0)}\mathrm{R}_1 + p_2^{(0)}\mathrm{R}_2 - p_1^{(0)}p_2^{(0)}\mathrm{R}_1\mathrm{R}_2 - \rho^{(1)}\sqrt{p_1^{(0)}\mathrm{R}_1 p_2^{(0)}\mathrm{R}_2(1-p_1^{(0)}\mathrm{R}_1)(1-p_2^{(0)}\mathrm{R}_2)}}{1 - q_1^{(0)}q_2^{(0)} - \rho^{(0)}\sqrt{p_1^{(0)}p_2^{(0)}q_1^{(0)}q_2^{(0)}}} \tag{A.3}$$

(iii) The odds ratio for the composite endpoint, OR_*, is determined by the six parameters $p_1^{(0)}$, $p_2^{(0)}$, OR_1, OR_2, $\rho^{(0)}$, $\rho^{(1)}$ and has the following expression:

$$\mathrm{OR}_* = \frac{\frac{\left(1+\frac{\mathrm{OR}_1 p_1^{(0)}}{1-p_1^{(0)}}\right)\left(1+\frac{\mathrm{OR}_2 p_2^{(0)}}{1-p_2^{(0)}}\right) - 1 - \rho^{(1)}\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho^{(1)}\sqrt{\frac{\mathrm{OR}_1\mathrm{OR}_2 p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}}{\frac{\left(1+\frac{p_1^{(0)}}{(1-p_1^{(0)})}\right)\cdot\left(1+\frac{p_2^{(0)}}{(1-p_2^{(0)})}\right) - 1 - \rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}{1+\rho^{(0)}\sqrt{\frac{p_1^{(0)}p_2^{(0)}}{(1-p_1^{(0)})(1-p_2^{(0)})}}}} \tag{A.4}$$

Proof (Proof of Theorem A.1). (i), (ii) The two expressions (A.2) and (A.3) follow in a straightforward manner after noting that:

$$p_*^{(i)} = 1 - q_1^{(i)}q_2^{(i)} - \rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}} = p_1^{(i)} + p_2^{(i)} - p_1^{(i)}p_2^{(i)} - \rho^{(i)}\sqrt{p_1^{(i)}p_2^{(i)}q_1^{(i)}q_2^{(i)}} \tag{A.5}$$

and taking into account $p_k^{(1)} = \delta_k + p_k^{(0)}$ and $p_k^{(1)} = p_k^{(0)}\mathrm{R}_1$.

(iii) Replacing the probabilities of the composite endpoint with its expression in terms of the marginal parameters plus the correlation (A.5), we have:

$$\mathrm{OR}_* = \left(\frac{1 - q_1^{(1)} q_2^{(1)} - \rho^{(1)} \sqrt{\frac{p_1^{(1)} p_2^{(1)}}{q_1^{(1)} q_2^{(1)}}}}{q_1^{(1)} q_2^{(1)} + \rho^{(1)} \sqrt{\frac{p_1^{(1)} p_2^{(1)}}{q_1^{(1)} q_2^{(1)}}}} \right) \cdot \left(\frac{1 - q_1^{(0)} q_2^{(0)} - \rho^{(0)} \sqrt{\frac{p_1^{(0)} p_2^{(0)}}{q_1^{(0)} q_2^{(0)}}}}{q_1^{(0)} q_2^{(0)} + \rho^{(0)} \sqrt{\frac{p_1^{(0)} p_2^{(0)}}{q_1^{(0)} q_2^{(0)}}}} \right)^{-1}$$

$$= \left(\frac{\frac{1}{q_1^{(1)} q_2^{(1)}} - 1 - \rho^{(1)} \sqrt{\frac{p_1^{(1)} p_2^{(1)}}{q_1^{(1)} q_2^{(1)}}}}{1 + \rho^{(1)} \sqrt{\frac{p_1^{(1)} p_2^{(1)}}{q_1^{(1)} q_2^{(1)}}}} \right) \left(\frac{\frac{1}{q_1^{(0)} q_2^{(0)}} - 1 - \rho^{(0)} \sqrt{\frac{p_1^{(0)} p_2^{(0)}}{q_1^{(0)} q_2^{(0)}}}}{1 + \rho^{(0)} \sqrt{\frac{p_1^{(0)} p_2^{(0)}}{q_1^{(0)} q_2^{(0)}}}} \right)^{-1}$$

Notice that:

$$\begin{aligned}
\frac{1}{q_1^{(i)} q_2^{(i)}} &= (1 + O_1^{(i)})(1 + O_2^{(i)}) \\
\frac{1}{q_1^{(1)} q_2^{(1)}} &= (1 + \mathrm{OR}_1 O_1^{(0)})(1 + \mathrm{OR}_2 O_2^{(0)}) \\
\frac{p_1^{(1)} p_2^{(1)}}{q_1^{(1)} q_2^{(1)}} &= \mathrm{OR}_1 \mathrm{OR}_2 O_1^{(0)} O_2^{(0)}
\end{aligned}$$

Hence:

$$\mathrm{OR}_* = \left(\frac{(1 + \mathrm{OR}_1 O_1^{(0)})(1 + \mathrm{OR}_2 O_2^{(0)}) - 1 - \rho^{(1)} \sqrt{\mathrm{OR}_1 \mathrm{OR}_2 O_1^{(0)} O_2^{(0)}}}{1 + \rho^{(1)} \sqrt{\mathrm{OR}_1 \mathrm{OR}_2 O_1^{(0)} O_2^{(0)}}} \right) \cdot$$

$$\cdot \left(\frac{(1 + O_1^{(0)})(1 + O_2^{(0)}) - 1 - \rho^{(0)} \sqrt{O_1^{(0)} O_2^{(0)}}}{1 + \rho^{(0)} \sqrt{O_1^{(0)} O_2^{(0)}}} \right)^{-1}$$

By replacing $O_k^{(0)}$ by $\frac{p_k^{(0)}}{1 - p_k^{(0)}}$, we obtain (A.4).

Theorem A.2 (Risk difference performance). *Assume that $p_k^{(0)} < 1/2$ and $\delta_k < 0$ $(k = 1, 2)$. We denote by $\delta_*(\rho, \theta, \lambda)$ the risk difference for the composite endpoint function described in* (A.2)*, specifically in terms of the vector of event rates θ, the marginal effects λ and the correlation ρ. Then, the risk difference for the composite endpoint for a given θ and λ is an increasing function with respect to ρ.*

Proof (Proof of Theorem A.2). Observe that the difference in proportions (A.2) can be written as:

$$\delta_*(\rho, \theta, \lambda) = x(\theta, \lambda) + \rho \cdot y(\theta, \lambda).$$

where $x(\theta, \lambda) = \delta_1 q_2^{(0)} + \delta_2 q_1^{(0)} - \delta_1 \delta_2$ and $y(\theta, \lambda) = \sqrt{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}} - \sqrt{p_1^{(1)} p_2^{(1)} q_1^{(1)} q_2^{(1)}}$. Then: $\delta_*(\rho+\epsilon; \theta) - \delta_*(\rho; \theta) = \epsilon \cdot y(\theta, \lambda)$. Therefore, $\delta_*(\rho; \theta)$ is an increasing function if and only if $y(\theta, \lambda) > 0$, $\forall \lambda, \theta$, which is equivalent to showing that:

$$\frac{p_1^{(0)} p_2^{(0)} q_1^{(0)} q_2^{(0)}}{p_1^{(1)} p_2^{(1)} q_1^{(1)} q_2^{(1)}} > 1$$

It is enough to prove that for $k = 1, 2$,

$$\frac{p_k^{(1)} q_k^{(1)}}{p_k^{(0)} q_k^{(0)}} < 1$$

To ease the notation call $p_k^{(1)} = a$ and $p_k^{(0)} = b$; and, by assuming $a < b < 1/2$, that implies $a - b < 0$ and $a + b < 1$. We need to prove that:

$$\frac{a(1-a)}{b(1-b)} = \frac{a - a^2}{b - b^2} < 1 \Leftrightarrow b - a < b^2 - a^2 = (b + a)(b - a)$$

Since $b - a > 0$ and $a + b < 1$, then we have $(a + b)(b - a) < (b - a)$. As a consequence $y(\theta, \lambda) > 0$ and the risk difference of the composite endpoint is an increasing function with respect to the correlation.

A.2 Derivation of the sample size for the composite binary endpoint

We establish the sample size formulae for the composite endpoint in terms of the margins and derive its properties, as outlined in Sections 2.4 and 2.6.

A.2.1 Sample size performance according to the correlation

Lemma A.1. *Let $N(p, d)$ denote the sample size function for testing the difference in proportions under the unpooled variance estimate, where p denotes the probability under the control group and d the relevant difference to be detected, that is:*

$$N(p,d) = \left(\frac{z_\alpha + z_\beta}{d}\right)^2 \cdot (p \cdot (1-p) + (p+d) \cdot (1-p-d)) \tag{A.6}$$

It follows that $N(p,d)$ is an increasing function with respect to p and with respect to d.

Proof. Observe that:

$$\frac{\partial}{\partial p} N\left(p,d\right) = \frac{\left(z_\alpha + z_\beta\right)^2 \left(2 - 4\,p - 2\,d\right)}{d^2}$$

Assuming $p < 0.5$, then $1-2p > 0$ and $2-4p0-2d > 0$. Therefore $\frac{\partial}{\partial p} N\left(p,d\right) > 0$, the sample size is increasing with respect to p. Moreover,

$$\begin{aligned}\frac{\partial}{\partial d} N(p,d) = -2\,\frac{\left(z_\alpha + z_\beta\right)^2 \left(p\left(1-p\right) + \left(d+p\right)\left(1-p-d\right)\right)}{d^3} \\ + \frac{\left(z_\alpha + z_\beta\right)^2 \left(1 - 2p - 2\,d\right)}{d^2}\end{aligned}$$

Note that $1 - 2p - 2d > 0$ and therefore, $\frac{\partial}{\partial d} N(p,d) > 0$; thus, the sample size is increasing with respect to d.

Theorem 1. Let θ and λ be the vectors of, respectively, marginal event rates and effect sizes for the composite components, and we denote by ρ the correlation between both components. Then, the sample size $n(\theta, \lambda, \rho)$, for a given θ and λ is an increasing function of the correlation ρ.

Proof (Proof of Theorem 1).

Since the probability of observing the composite event is given by θ and ρ (see equation (A.5)), and the risk difference for the composite endpoint is given by λ, θ and ρ (see equation (A.2)), then the sample size for the composite endpoint computed by $n(\theta, \lambda, \rho) = N\left(p_*(\theta,\rho), \delta_*(\lambda,\theta,\rho)\right)$ is a function of λ, θ and ρ.

To prove that the sample size for the composite endpoint $N\left(p_*(\theta,\rho), \delta_*(\lambda,\theta,\rho)\right)$ increases with ρ, we will show that:

$$\begin{aligned}\frac{\partial N(p_*(\theta,\rho), \delta_*(\lambda,\theta,\rho))}{\partial \rho} = \frac{\partial N(p_*(\theta,\rho), \delta_*(\lambda,\theta,\rho))}{\partial p_*(\theta,\rho)} \cdot \frac{\partial p_*(\theta,\rho)}{\partial \rho} + \\ + \frac{\partial N(p_*(\theta,\rho), \delta_*(\lambda,\theta,\rho))}{\partial \delta_*(\lambda,\theta,\rho)} \cdot \frac{\partial \delta_*(\lambda,\theta,\rho)}{\partial \rho} > 0\end{aligned} \tag{A.7}$$

From now on we will omit θ, λ and ρ and use p_* and δ_* instead of $p_*(\theta,\rho)$ and $\delta_*(\lambda,\theta,\rho)$. From Lemma A.1, the sample size $N(p,d)$ in (A.6) is increasing with respect to the treatment effect, d, and with respect to the probability of observing the event under control group, p, hence:

$$\frac{\partial N(p_*,\delta_*)}{\partial p_*} \;>\; 0 \quad \text{and} \quad \frac{\partial N(p_*,\delta_*)}{\partial \delta_*} \;>\; 0$$

We denote by:

$$\frac{\partial p_*(\theta,\rho)}{\partial \rho} \;=\; -a \quad \text{and} \quad \frac{\partial \delta_*(\rho)}{\partial \rho} \;=\; a-b$$

where $a, b > 0$ and, from Theorem A.2, $a - b > 0$. Then we have:

$$\frac{\partial N(p_*,\delta_*)}{\partial \rho} = (a-b)\cdot\left(-2\,\frac{(z_\alpha+z_\beta)^2\,(p_*\;(1-p_*(\theta,\rho))+(\delta_*+p_*)\,(1-p_*-\delta_*))}{{\delta_*}^3} + \frac{(z_\alpha+z_\beta)^2\,(1-2\,p_*-2\delta_*)}{{\delta_*}^2}\right) - a\cdot\frac{(z_\alpha+z_\beta)^2\,(2-4\,p_*-2\delta_*)}{{\delta_*}^2}$$

and this is positive if and only if:

$$(a-b)\big(-2\frac{p_*\;(1-p_*)+(\delta_*+p_*)\,(1-p_*-\delta_*)}{\delta_*}+1-2(p_*+\delta_*)\big)-(2-4\,p_*-2\delta_*)\,a \tag{A.8}$$

(A.8) is positive. Then we have:

$$\begin{aligned}&-2\frac{(a-b)}{\delta_*}\,(p_*\,(1-p_*)+(\delta_*+p_*)\,(1-p_*-d))-b\,(1-2\,p_*-2\delta_*)\;+a\,(-1+2\,p_*)\\ &>-2\frac{(a-b)}{\delta_*}p_*\,(1-p_*)-2\frac{(a-b)}{\delta_*}\,(\delta_*+p_*)\,(1-p_*-\delta_*)-2a\,(1-p_*-\delta_*)+2ap_*\end{aligned} \tag{A.9}$$

Then (A.8) > (A.9), because $a > b$. Note that the first and forth terms are positive, so we end if we see that the second plus third are also positive. This follows from the fact that:

$$-2\frac{(a-b)}{\delta_*}\,(\delta_*+p_*)-2a>0\Leftrightarrow a\left(1+\frac{\delta_*+p_*}{\delta_*}\right)<b\left(\frac{\delta_*+p_*}{\delta_*}\right)\Leftrightarrow\frac{a}{b}>\frac{\frac{\delta_*+p_*}{\delta_*}}{1+\frac{\delta_*+p_*}{\delta_*}}$$

Since $a, b > 0$ and $a - b > 0$, we have $\frac{a}{b} > 1$; and since $\left(\frac{\delta_*+p_*}{\delta_*}\right)$, $\left(1+\frac{\delta_*+p_*}{\delta_*}\right) < 0$, we have $\left(\frac{\delta_*+p_*}{\delta_*}\right) / \left(1+\frac{\delta_*+p_*}{\delta_*}\right) \in (0,1)$. Therefore (A.7) is positive, as we intended to prove.

Appendix B
Appendix of Bofill and Gómez (2018)

B.1 Additional tables and figures

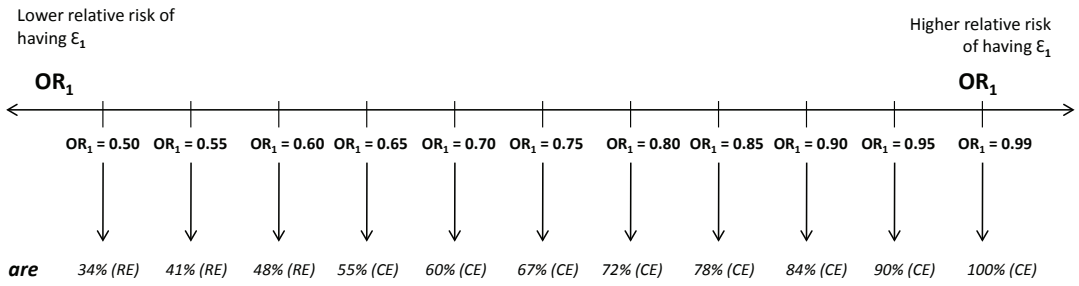


Fig. B.1 Percentage of scenarios in which the composite endpoint should be used depending on OR_1.

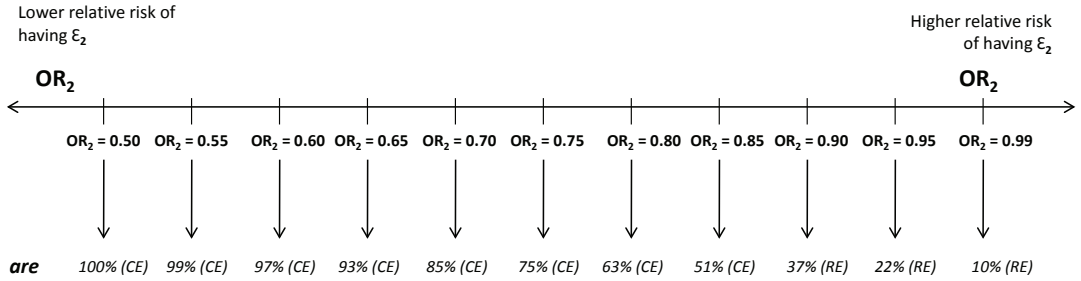


Fig. B.2 Percentage of scenarios in which the composite endpoint should be used depending on OR_2.

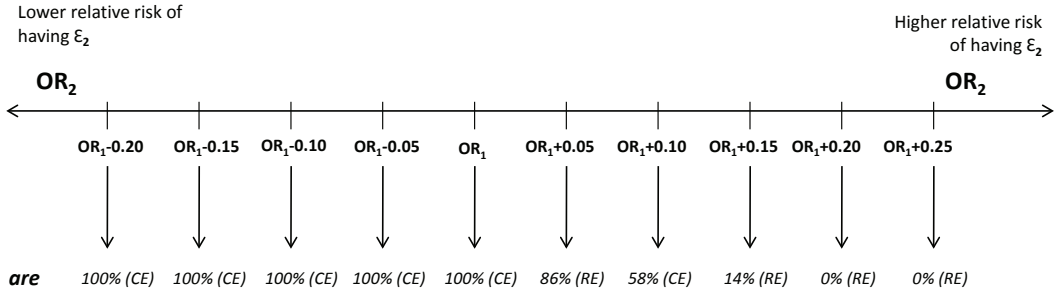


Fig. B.3 Percentage of scenarios in which the composite endpoint should be used depending on OR_2 when $OR_1 = 0.7$.

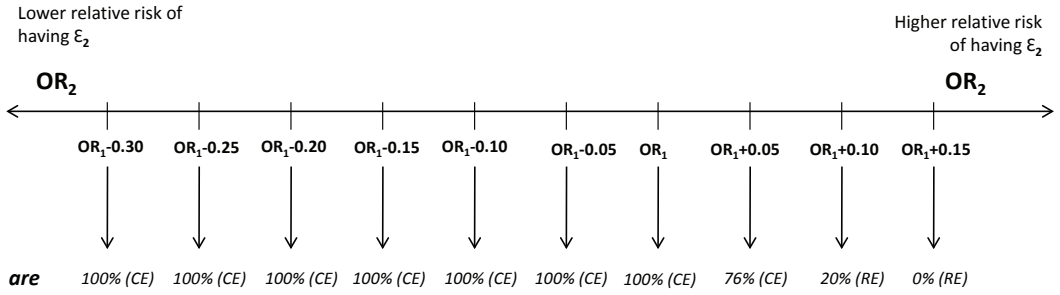


Fig. B.4 Percentage of scenarios in which the composite endpoint should be used depending on OR_2 when $OR_1 = 0.8$.

Table B.1 Recommendations in terms of treatment effects of the relevant and the additional endpoint, large ($0.5 \leq$ OR < 0.7), medium ($0.7 \leq$ OR < 0.9) or low ($0.9 \leq$ OR < 1). Each cell indicates whether the relevant endpoint (RE) ($are \leq 1.1$) or composite endpoint (CE) ($are > 1.1$) should be used and, in parentheses, the percentage of cases in which composite is preferred based on the scenarios described in Table 3.2.

	Large treatment effect on ε_2	Medium treatment effect on ε_2	Low treatment effect ε_2
Large treatment effect on ε_1	CE (80.97%)	RE (15.65%)	RE (0.00%)
Medium treatment effect on ε_1	CE (99.84%)	CE (74.53%)	RE (4.23%)
Low treatment effect ε_1	CE (100.00%)	CE (99.99%)	CE (63.89%)

Table B.2 Recommendations in terms of degree of association between endpoints, weak ($0 < \rho < 0.3$), medium-weak ($0.3 \leq \rho < 0.6$), medium-strong ($0.6 \leq \rho < 0.8$), strong ($0.8 \leq \rho < 1$); treatment effects of the relevant and the additional endpoint, large ($0.5 \leq \text{OR} < 0.7$), medium ($0.7 \leq \text{OR} < 0.9$) or low ($0.9 \leq \text{OR} < 1$); event rates in control group for the relevant and additional endpoints, low ($p \leq 0.025$), medium-low ($0.025 \leq p \leq 0.05$), medium-large ($0.05 \leq p \leq 0.075$), large ($\leq p > 0.075$). Each cell indicates whether the relevant endpoint (RE) ($are \leq 1.1$) or composite endpoint (CE) ($are > 1.1$) should be used and, in parentheses, the percentage of cases in which composite is preferred based on the scenarios described in Table 3.2.

Correlation	Weak	Medium-weak	Medium-strong	Strong
Large treatment effect on ε_2	CE (97.29%)	CE (93.35%)	CE (87.94%)	CE (76.10%)
Medium treatment effect on ε_2	CE (68.08%)	CE (60.14%)	CE/RE (53.33%)	CE/RE (50.49%)
Low treatment effect ε_2	RE (21.53%)	RE (19.38%)	RE (18.93%)	RE (22.12%)
Large treatment effect on ε_1	RE (43.35%)	RE (35.82%)	RE (29.52%)	RE (28.45%)
Medium treatment effect on ε_1	CE (69.67%)	CE (64.59%)	CE (59.13%)	CE/RE (51.88%)
Low treatment effect ε_1	CE (91.33%)	CE (89.94%)	CE (88.09%)	CE (80.57%)
Low event rate for ε_1	CE (73.08%)	CE (71.92%)	CE (65.38%)	CE (58.18%)
Medium-low event rate for ε_1	CE (67.49%)	CE (63.06%)	CE (62.62%)	CE (56.39%)
Medium-large event rate for ε_1	CE (63.84%)	CE (58.08%)	CE/RE (54.85%)	CE (55.48%)
Large event rate for ε_1	CE (61.03%)	CE/RE (54.54%)	RE (48.06%)	RE (42.09%)
Low event rate for ε_2	CE (58.35%)	RE (49.03%)	RE (47.41%)	RE (43.03%)
Medium-low event rate for ε_2	CE (65.63%)	CE (59.95%)	CE/RE (50.24%)	RE (46.93%)
Medium-large event rate for ε_2	CE (68.45%)	CE (64.58%)	CE (58.19%)	RE (47.65%)
Large event rate for ε_2	CE (70.06%)	CE (66.95%)	CE (64.29%)	CE (61.53%)

Appendix C
Appendix of Bofill and Gómez (2020)

Define, for the i-th group ($i = 0, 1$), the counting processes $N_{ij}(t) = I\{T_{ij} \wedge C_{ij} \leq t, \delta_{ij} = 1\}$, $\bar{N}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} N_{ij}(t)$; the at-risk processes $Y_{ij}(t) = I\{T_{ij} \wedge C_{ij} \geq t\}$, $\bar{Y}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} Y_{ij}(t)$; and let $\Lambda^{(i)}(t) = -\log\{S^{(i)}(t)\}$ be the cumulative hazard function. Then, $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(s) d\Lambda^{(i)}(s)$ is a zero-mean martingale with respect to the filtration $\mathcal{F}(t) = \sigma < \{N_{ij}(t), Y_{ij}(t)\}, j = 1, ..., n^{(i)}, i = 0, 1 >$. Consider $\bar{M}^{(i)}(t) = \sum_{j=1}^{n^{(i)}} M_{ij}(t)$. We write $dN(t)$ for the increment $N((t+dt)-) - N(t-)$ of the counting process $N(\cdot)$ over the small time interval $[t, t + dt)$.

Consider the subgroup of patients who had the binary outcome before the time-to-event outcome, from now on, we call them responders subgroup. Let $S_X^{(i)}(t) = \mathrm{P}(T_{ij} > t | X_{ij} = 1)$, $\Lambda_X^{(i)}(t)$ and $\lambda_X^{(i)}(t)$ be the survival, cumulative hazard and hazard functions, respectively, for the responders. Define:

$$N_{ij,X}(t) = I\{T_{ij} \wedge C_{ij} \leq t, \delta_{ij} = 1, X_{ij} = 1\} = N_{ij}(t) \cdot X_{ij} \tag{C.1}$$

$$Y_{ij,X}(t) = I\{T_{ij} \wedge C_{ij} \geq t, X_{ij} = 1\} = Y_{ij}(t) \cdot X_{ij} \tag{C.2}$$

and $\bar{N}_X^{(i)}(t) = \sum_{j=1}^{n_i} N_{ij,X}(t)$ and $\bar{Y}_X^{(i)}(t) = \sum_{j=1}^{n_i} Y_{ij,X}(t)$. Then, $M_{ij,X}(t) = N_{ij,X}(t) - \int_0^t Y_{ij,X}(s) d\Lambda_X^{(i)}(s)$ is a zero-mean martingale with respect to the filtration $\mathcal{F}(t) = \sigma < \{N_{ij,X}(t), Y_{ij,X}(t)\}, j = 1, ..., n^{(i)}, i = 0, 1 >$. Consider $\bar{M}_X^{(i)}(t) = \sum_{j=1}^{n_i} M_{ij,X}(t)$.

A sequence of random vectors X_n that converges in probability to X as $n \to +\infty$ will be denoted by $X_n \xrightarrow{p} X$. The convergence in distribution will be written as $X_n \xrightarrow{d} X$.

C.1 Main results and their proofs

Lemma 4.1:
Let $\mathbf{U}_n^{\omega}(Q)$ be the statistic defined by:

$$\mathbf{U}_n^{\omega}(Q) = \omega_b \cdot \frac{U_{b,n}}{\hat{\sigma}_b} + \omega_s \cdot \frac{U_{s,n}(Q)}{\hat{\sigma}_s}$$

where $U_{b,n}$ is the statistic given in (4.3) and $U_{s,n}(Q)$ is the statistic given in (4.4) with $\hat{Q}(t)$ replaced by $Q(t)$, that is:

$$U_{s,n}(Q) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\int_{\tau_0}^{\tau} Q(t) \cdot \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt \right)$$

for some real numbers $\omega_b, \omega_s \in (0,1)$, such that $\omega_b + \omega_s = 1$, and for a function $Q(\cdot)$ satisfying the conditions outlined in Section 4.3.1. Then, the $\mathcal{L}$-statistic $\mathbf{U}_n^{\omega}(\hat{Q})$, given in (4.2), can be written as:

$$\mathbf{U}_n^{\omega}(\hat{Q}) = \mathbf{U}_n^{\omega}(Q) + \omega_s \cdot \frac{E_n}{\hat{\sigma}_s}$$

where

$$E_n = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot (\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t)) dt$$

converges in probability to 0. Hence, the asymptotic distribution of the statistic $\mathbf{U}_n^{\omega}(Q)$ is the same as that of $\mathbf{U}_n^{\omega}(\hat{Q})$.

Proof of Lemma 4.1:
The proof is a direct consequence of the asymptotic representation of the time-to-event statistic $U_{s,n}(\hat{Q})$ which can be written as $U_{s,n}(Q) + E_n$, where:

$$U_{s,n}(Q) = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\int_{\tau_0}^{\tau} Q(t) \cdot \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt \right)$$

and

$$E_n = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \left(\int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt \right)$$

and where E_n is a second-order term that is asymptotically negligible. For a detailed proof of this representation, we refer to (Gu et al., 1999).

In order to prove that $\mathbf{U}_n^{\omega}(\hat{Q})$ and $\mathbf{U}_n^{\omega}(Q)$ are asymptotically equivalent, we first note that:

$$\mathbf{U}_n^{\omega}(\hat{Q}) - \mathbf{U}_n^{\omega}(Q) = \omega_s \cdot \frac{E_n}{\hat{\sigma}_s}$$

Since

$$E_n = \sqrt{\frac{n^{(0)}n^{(1)}}{n}} \sum_{i=0,1} (-1)^{i+1} \left(\int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot (\hat{S}^{(i)}(t) - S^{(i)}(t)) dt \right.$$
$$\left. + \int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot S^{(i)}(t) dt \right),$$

following Lemma 1 from Gu et al. (1999), for $i = 0, 1$, we have:

$$\sqrt{\frac{n^{(0)}n^{(1)}}{n}} \int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot (\hat{S}^{(i)}(t) - S^{(i)}(t)) dt \xrightarrow{p} 0$$

and

$$\sqrt{\frac{n^{(0)}n^{(1)}}{n}} \int_{\tau_0}^{\tau} (\hat{Q}(t) - Q(t)) \cdot S^{(i)}(t) dt \xrightarrow{p} 0$$

because $\hat{Q}(t) - Q(t) \xrightarrow{p} 0$ uniformly for t. Thus, $E_n \xrightarrow{p} 0$ and since $\hat{\sigma}_s^2 \xrightarrow{p} \sigma_s^2$ and $\hat{\sigma}_s^2$ is bounded away from 0, we conclude that $\mathbf{U}_n^{\omega}(\hat{Q}) - \mathbf{U}_n^{\omega}(Q) \xrightarrow{p} 0$. Hence, by Slutsky's Theorem, both statistics are asymptotically equivalent:

$$\mathbf{U}_n^{\omega}(\hat{Q}) = \left(\mathbf{U}_n^{\omega}(\hat{Q}) - \mathbf{U}_n^{\omega}(Q) \right) + \mathbf{U}_n^{\omega}(Q) \xrightarrow{d} \mathbf{U}_n^{\omega}(Q)$$

□

Theorem 4.1:
Let $\mathbf{U}_n^{\omega}(\hat{Q})$ be the statistic defined in (4.2). Under the conditions outlined in 4.3.1, if the null hypothesis $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$ holds, $\mathbf{U}_n^{\omega}(\hat{Q})$ converges in distribution, as $n \to +\infty$, to a normal distribution as follows:

$$\mathbf{U}_n^{\omega}(\hat{Q}) \to N\left(0, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \cdot \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

where σ_b^2, σ_s^2 stand for the variances of $U_{b,n}$ and $U_{s,n}(Q)$, respectively, and σ_{bs} is the covariance between $U_{b,n}$ and $U_{s,n}(Q)$. Their corresponding expressions are given by:

$$\sigma_b^2 = \sum_{i=0,1} \left(1-\pi^{(i)}\right) p^{(i)}\left(\tau_b\right)\left(1-p^{(i)}\left(\tau_b\right)\right)$$

$$\sigma_s^2 = -\sum_{i=0,1}\left(1-\pi^{(i)}\right)\int_{\tau_0}^{\tau}\frac{(K_\tau^{(i)}(t))^2}{(S^{(i)}(t))^2 G^{(i)}(t)} dS^{(i)}(t)$$

$$\sigma_{bs} = \sum_{i=0,1}\left(1-\pi^{(i)}\right)\cdot\left(I\{\tau_{\max}=\tau_b\}\cdot\int_{\tau_0}^{\tau_b}\frac{K_{\tau_b}^{(i)}(t)}{S^{(i)}(t)}\cdot\left(p_N^{(i)}(t)-p^{(i)}\left(\tau_b\right)\right)dS^{(i)}(t)\right.$$
$$\left.+\int_{\tau_{\max}}^{\tau}\frac{K_\tau^{(i)}(t)}{S^{(i)}(t)}\cdot p^{(i)}\left(\tau_b\right)\left(dS_X^{(i)}(t)-dS^{(i)}(t)\right)\right)$$

where $\tau_{\max}=\max(\tau_0,\tau_b)$, $K_{\tau_*}^{(i)}(t)=\int_t^{\tau_*}Q(u)S^{(i)}(u)du$ ($\tau_*=\tau$ or τ_b), $p_N^{(i)}(t)=P(X_{ij}=1|dN_{ij}(t)=1)$, and $S_X^{(i)}(t)=P(T_{ij}>t|X_{ij}=1)$ for $i=0,1$.

Proof of Theorem 4.1:
Since $\mathbf{U}_n^\omega(\hat{Q})-\mathbf{U}_n^\omega(Q)\xrightarrow{p}0$, it is sufficient to find the asymptotic distribution of $\mathbf{U}_n^\omega(Q)$. Suppose that the global null hypothesis $\mathrm{H}_0:\mathrm{H}_{s,0}\cap\mathrm{H}_{b,0}$ holds. The statistic $\mathbf{U}_n^\omega(Q)$ is a weighted sum of two statistics: $U_{b,n}/\hat{\sigma}_b$ for the sub-hypotheses $\mathrm{H}_{b,0}$, and $U_{s,n}(Q)/\hat{\sigma}_s$ for the sub-hypotheses $\mathrm{H}_{s,0}$. Under $\mathrm{H}_{b,0}$, the statistic $U_{b,n}$ asymptotically follows a 0-mean normal distribution with variance σ_b^2 (Lachin, 1981), where σ_b^2 is given by (4.5). On the other hand, the statistic $U_{s,n}(Q)$ asymptotically follows a 0-mean normal distribution with variance σ_s^2 under $\mathrm{H}_{s,0}$ (Gu et al., 1999), where σ_s^2 is given by (4.6). Since $\hat{\sigma}_b$ and $\hat{\sigma}_s$, given in (4.10) and (4.11), consistently estimate σ_b^2 and σ_s^2, by Slutsky's Theorem, we have that both $U_{b,n}/\hat{\sigma}_b$ and $U_{s,n}(Q)/\hat{\sigma}_s$ are asymptotically $N(0,1)$ under H_0. Then, we have that, as $n\to+\infty$, the asymptotic distribution of $\mathbf{U}_n^\omega(Q)$ under H_0 is:

$$\mathbf{U}_n^\omega(Q)\xrightarrow{d}N\left(0,\omega_b^2+\omega_s^2+2\omega_b\omega_s\frac{\sigma_{bs}}{\sigma_b\sigma_s}\right)$$

where σ_{bs} denotes the covariance between $U_{b,n}$ and $U_{s,n}(Q)$. The expression of the covariance and its corresponding derivation are postponed to Appendix C.2.1.

□

Theorem 4.2:
Let $\mathbf{U}_n^\omega(\hat{Q})$ be the statistic defined in (2). Under the conditions outlined in 3.1., consider the following sequences of contiguous alternatives for both binary and time-to-event hypotheses satisfying, as $n\to+\infty$:

$$\sqrt{n}(p_n^{(1)} - p^{(0)}) \to g$$

and

$$\sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) \to \mathcal{G}(t)$$

for some constant $g \in \mathbb{R}^+$ and bounded function $\mathcal{G}(\cdot) \in \mathbb{R}^+$, and $\forall t \in [\tau_0, \tau]$. Then, under contiguous alternatives of the form:

$$\mathrm{H}_{1,n} : \sqrt{n}(p_n^{(1)} - p^{(0)}) = g \ \text{ and } \ \sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) = \mathcal{G}(t), \quad \forall t \in [\tau_0, \tau]$$

we have that:

$$\mathbf{U}_n^{\omega}(\hat{Q}) \to N\left(\omega_b g + \omega_s \int_{\tau_0}^{\tau} Q(t)\mathcal{G}(t)dt, \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\sigma_{bs}}{\sigma_b \cdot \sigma_s}\right)$$

in distribution as $n \to +\infty$, where σ_b^2, σ_s^2 and σ_{bs} are given in (4.5), (4.6) and (4.7), respectively.

Proof of Theorem 4.2:
Suppose $\mathrm{H}_{1,n} : \sqrt{n}(p_n^{(1)} - p^{(0)}) = g$ and $\sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) = \mathcal{G}(t)$, $\forall t \in [\tau_0, \tau]$ holds. Let us consider the sub-hypotheses $\mathrm{H}_{1b,n} : \sqrt{n}(p_n^{(1)} - p^{(0)}) = g$ and $\mathrm{H}_{1s,n} : \sqrt{n}(S_n^{(1)}(t) - S^{(0)}(t)) = \mathcal{G}(t)$, $\forall t \in [\tau_0, \tau]$. Analogously to the proof of Theorem 4.2, the statistics $U_{b,n}/\hat{\sigma}_b$ and $U_{s,n}(Q)/\hat{\sigma}_s$ follow, respectively, a normal distribution with mean g and variance equal to 1 under $\mathrm{H}_{b,0}$ and mean $\int_0^{\tau} Q(t)\mathcal{G}(t)dt$ and variance equal to 1 under $\mathrm{H}_{1s,n}$ Gu et al. (1999). Therefore, we have that $\mathbf{U}_n^{\omega}(Q) \xrightarrow{d} N(\mu_c,\ \sigma_c^2)$, where $\mu_c = \omega_b g + \omega_s \int_{\tau_0}^{\tau} Q(t)\mathcal{G}(t)dt$, and $\sigma_c^2 = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\sigma_{bs}}{\sigma_b\sigma_s}$. □

Theorem 4.3:
Let $\mathbf{U}_n^{\omega}(\hat{Q})$ be the statistic defined in (4.2), and let σ_b^2, σ_s^2 and σ_{bs} be the variances and covariance given in (4.5), (4.6) and (4.7), respectively. The asymptotic variance of $\mathbf{U}_n^{\omega}(\hat{Q})$, given in Theorem 4.2, can be consistently estimated by:

$$\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q})) = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\hat{\sigma}_{bs}}{\hat{\sigma}_b \cdot \hat{\sigma}_s}$$

where $\hat{\sigma}_b$, $\hat{\sigma}_s$, and $\hat{\sigma}_{bs}$ denote the estimates of σ_b, σ_s and σ_{bs}, and are given by:

$$\hat{\sigma}_b^2 = \hat{p}\,(\tau_b)\,(1-\hat{p}\,(\tau_b))$$

$$\hat{\sigma}_s^2 = -\int_{\tau_0}^{\tau} \frac{(\hat{K}_\tau(t))^2}{\hat{S}(t)\hat{S}(t-)} \cdot \frac{n^{(0)}\hat{G}^{(0)}(t-) + n^{(1)}\hat{G}^{(1)}(t-)}{\hat{G}^{(0)}(t-)\hat{G}^{(1)}(t-)} d\hat{S}(t)$$

$$\hat{\sigma}_{bs} = -\int_{\tau_0}^{\tau_b} \hat{K}_{\tau_b}(t) \left(\sum_{i=0,1} \frac{n-n^{(i)}}{n} \cdot \hat{\lambda}_{X,T}^{(i)}(t)dt + \frac{\hat{p}\,(\tau_b)\cdot d\hat{S}(t)}{\hat{S}(t)} \right)$$
$$+ \int_{\tau_b}^{\tau} \frac{\hat{K}_\tau(t)\cdot \hat{p}\,(\tau_b)}{\hat{S}(t-)} \left(-\frac{\hat{S}(t-)\cdot d\hat{S}(t)}{\hat{S}(t)} + \sum_{i=0,1} \frac{n-n^{(i)}}{n} \cdot \frac{\hat{S}_X^{(i)}(t-)\cdot d\hat{S}_X^{(i)}(t)}{\hat{S}_X^{(i)}(t)} \right)$$

where $\hat{K}_{\tau_*}(t) = \int_t^{\tau_*} \hat{Q}(u)\hat{S}(u)du$ ($\tau_* = \tau$ or τ_b), $\hat{S}_X^{(i)}(t)$ is the Kaplan-Meier estimator of $S_X^{(i)}(t)$; and $\hat{\lambda}_{X,T}^{(i)}(t)$ is the estimator of $\lambda_{X,T}^{(i)}(t) = \lim_{dt\to 0} P(X_{ij}=1, t \leq T_{ij} < t+dt | T_{ij} > t)/dt$. Kernel-density methods are used in the estimation of $\lambda_{X,T}^{(i)}(t)$.

Proof of Theorem 4.3:
As stated in Theorem 4.1, the asymptotic variance of $\mathbf{U}_n^\omega(\hat{Q})$ is given by:

$$\mathrm{Var}(\mathbf{U}_n^\omega(\hat{Q})) = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s \frac{\sigma_{bs}}{\sigma_b\sigma_s} \tag{C.3}$$

This variance can be consistently estimated by using the plug-in method, i.e., by substituting $\hat{\sigma}_b^2$, $\hat{\sigma}_s^2$, and $\hat{\sigma}_{bs}$ for σ_b^2, σ_s^2 and σ_{bs}, respectively, and $\hat{Q}(\cdot)$ for $Q(\cdot)$. Consistent estimators for σ_b^2, σ_s^2 are given by (4.10) and (4.11). For further details of these estimators, we refer to (Lachin, 1981) and (Pepe and Fleming, 1989).

On the other hand, the covariance can be estimated by (4.12). The differences between the theoretical and estimated covariance expressions ((4.7) and (4.12)) arise from the fact that $p_N(t)$ can be expressed as the ratio of $\lambda_{X,T}^{(i)}(t)$ and the time-to-event hazard function $\lambda^{(i)}(s)ds = -\frac{dS^{(i)}(s)}{S^{(i)}(s)}$. Further details of the derivation and estimation of the covariance are postponed to Appendixes C.2.1 and C.2.2.

□

Theorem 4.4:
Let $\mathbf{U}_n^\omega(\hat{Q})$ be the statistic defined in (4.2), and let $\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))$ be the variance estimator given in (4.9). Consider the global null hypothesis H_0 ((4.1)) and let the normalized statistic of $\mathbf{U}_n^\omega(\hat{Q})$ be:

$$\mathbf{U}_n^\omega(\hat{Q}) \Big/ \sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^\omega(\hat{Q}))} \tag{C.4}$$

Then, the statistic defined in (C.4) converges in distribution to a standard normal distribution. Moreover, for positive $Q(\cdot)$, the statistic is consistent against any alternative hypothesis of the form of H_1 in (4.1) which contemplate differences and stochastic ordering alternatives for the binary and time-to-event outcomes, respectively.

Proof of Theorem 4.4:
Following the asymptotic results, under H_0, in Theorems 3.2 and 3.4, and relying on Slutsky's theorem, the statistic $\mathbf{U}_n^{\omega}(\hat{Q})/\sqrt{\widehat{\mathrm{Var}}(\mathbf{U}_n^{\omega}(\hat{Q}))}$ asymptotically follows a standard normal distribution. Moreover, since both $\sqrt{\frac{n^{(0)}n^{(1)}}{n}}U_{s,n}(\hat{Q})/\hat{\sigma}_s$ and $\sqrt{\frac{n^{(0)}n^{(1)}}{n}}U_{b,n}/\hat{\sigma}_b$ are consistent Pepe and Fleming (1991), the normalized statistic of $\mathbf{U}_n^{\omega}(\hat{Q})$ for $Q(\cdot) > 0$ is consistent against any alternative hypotheses of the form of H_1 in (4.1).

Theorem C.1. *Let $\boldsymbol{U}_n^{\omega}(\hat{Q})$ be the statistic defined in* (4.2)*, and let σ_b^2, σ_s^2 and σ_{bs}^2 be the variances and covariance given in* (4.5), (4.6) *and* (4.7)*, respectively. An unpooled estimate of the asymptotic variance of $\boldsymbol{U}_n^{\omega}(\hat{Q})$, given in Theorem 4.2, is:*

$$\widehat{\mathrm{Var}}_{up}(\boldsymbol{U}_n^{\omega}(\hat{Q})) = \omega_b^2 + \omega_s^2 + 2\omega_b\omega_s\frac{\hat{\sigma}_{bs,up}}{\hat{\sigma}_{b,up}\hat{\sigma}_{s,cup}} \tag{C.5}$$

where $\hat{\sigma}_{b,up}$, $\hat{\sigma}_{s,up}$, and $\hat{\sigma}_{bs,up}$ denote the unpooled estimates of σ_b, σ_s and σ_{bs}, given by:

$$\hat{\sigma}_{b,up}^2 = \sum_{i=0,1}\frac{n-n^{(i)}}{n}\cdot\hat{p}^{(i)}(\tau_b)\left(1-\hat{p}^{(i)}(\tau_b)\right) \tag{C.6}$$

$$\hat{\sigma}_{s,up}^2 = -\sum_{i=0,1}\frac{n-n^{(i)}}{n}\int_{\tau_0}^{\tau}\frac{(\hat{K}^{(i)}(t))^2}{\hat{S}^{(i)}(t)\hat{S}^{(i)}(t-)\hat{G}^{(i)}(t-)}d\hat{S}^{(i)}(t) \tag{C.7}$$

$$\begin{aligned}\hat{\sigma}_{bs,up} = \sum_{i=0,1}\frac{n-n^{(i)}}{n}\Bigg(&-\int_{\tau_0}^{\tau_b}\hat{K}_{\tau_b}^{(i)}(t)\left(\hat{\lambda}_{X,T}^{(i)}(t)dt+\frac{\hat{p}^{(i)}d\hat{S}^{(i)}(t)}{\hat{S}^{(i)}(t)}\right)\\&+\int_{\tau_b}^{\tau}\frac{\hat{K}_{\tau}^{(i)}(t)\hat{p}^{(i)}}{\hat{S}^{(i)}(t-)}\left(\frac{\hat{S}_X^{(i)}(t-)\cdot d\hat{S}_X^{(i)}(t)}{\hat{S}_X^{(i)}(t)}-\frac{\hat{S}^{(i)}(t-)\cdot d\hat{S}^{(i)}(t)}{\hat{S}^{(i)}(t)}\right)\Bigg)\end{aligned} \tag{C.8}$$

where $\hat{K}^{(i)}_{\tau_*}(t) = \int_t^{\tau_*} \hat{Q}(u)\hat{S}^{(i)}(u)du$ *(*$\tau_* = \tau$ *or* τ_b*),* $\hat{S}^{(i)}_X(t)$ *is the Kaplan-Meier estimator of* $S^{(i)}_X(t)$*; and* $\hat{\lambda}^{(i)}_{X,T}(t)$ *is the estimator of* $\lambda^{(i)}_{X,T}(t) = \lim_{dt\to 0} P(X_{ij} = 1, t \leq T_{ij} < t + dt | T_{ij} > t)/dt$ *and is estimated with kernel-based methods.*

C.2 Covariance derivation and estimation

C.2.1 Covariance derivation

Lemma C.1. *Assume that* $S^{(i)}(\tau) > 0$*,* $S^{(i)}_X(\tau) > 0$*, and* $G^{(i)}(\tau) > 0$ *for* $i = 0, 1$*. Then,*

$$\sup_{t\in[0,\tau]} |\bar{Y}^{(i)}(t)/n^{(i)} - y^{(i)}(t)| \xrightarrow{p} 0 \qquad \textit{and} \qquad \sup_{t\in[0,\tau]} |\bar{Y}^{(i)}_X(t)/n^{(i)} - y^{(i)}_X(t)| \xrightarrow{p} 0$$

where $y^{(i)}(t) = S^{(i)}(t)G^{(i)}(t)$ *and* $y^{(i)}_X(t) = p^{(i)}S^{(i)}_X(t)G^{(i)}(t)$*.*

Proof of Lemma A.
The proof follows analogous arguments as those used in the proof of Lemma 1 in the Appendix of Elashoff et al. (2012). The proof for both processes $\bar{Y}^{(i)}(t)/n^{(i)}$ and $\bar{Y}^{(i)}_X(t)/n^{(i)}$ mainly takes into account that at a given t,

$$\mathrm{E}(Y_{ij}(t)) = \mathrm{P}(Y_{ij}(t) = 1) = S^{(i)}(t-)G^{(i)}(t-) = S^{(i)}(t)G^{(i)}(t)$$

because $S^{(i)}(\cdot)$ and $G^{(i)}(\cdot)$ are continuous functions, and

$$\begin{aligned}\mathrm{E}(Y_{ij,X}(t)) &= \mathrm{P}(Y_{ij}(t) \cdot X_{ij} = 1) = \mathrm{P}(X_{ij} = 1)\mathrm{P}(Y_{ij}(t)|X_{ij} = 1) \\ &= p^{(i)}S^{(i)}_X(t-)C^{(i)}_X(t-)\end{aligned}$$

where $C^{(i)}_X(t-)$ is the censoring survival function for the responders. Since the binary outcome and censoring are independent, we have $C^{(i)}_X(t) = G^{(i)}(t)$ and because $S^{(i)}_X(\cdot)$ and $C^{(i)}_X(\cdot)$ are continuous functions it follows that $\mathrm{E}(Y_{ij,X}(t)) = p^{(i)}S^{(i)}_X(t)G^{(i)}(t) = y^{(i)}_X(t)$. The remainder of the proof is based on applying the weak law of large numbers for each t and the Lebesgue dominated convergence theorem to both $\bar{Y}^{(i)}(t)/n^{(i)}$ and $\bar{Y}^{(i)}_X(t)/n^{(i)}$.

Theorem C.2. *Assume that* $S^{(i)}(\tau) > 0$, $S_X^{(i)}(\tau) > 0$, $G^{(i)}(\tau) > 0$ *and that* $y^{(i)}(\tau) > 0$ *for* $i = 0, 1$. *Let* $U_{b,n}(\tau_b)$ *and* $U_{s,n}(\tau_0, \tau; \hat{Q})$ *be the binary and time-to-event statistics given in (3) and (4), respectively. Let* σ_{bs} *be the asymptotic covariance between* $U_{b,n}(\tau_b)$ *and* $U_{s,n}(\tau_0, \tau; \hat{Q})$ *under the null hypothesis* $\mathrm{H}_0 : \mathrm{H}_{s,0} \cap \mathrm{H}_{b,0}$. *Then:*

$$\sigma_{bs} = \sum_{i=0,1} \left(1 - \pi^{(i)}\right) \left(\int_{\tau_0}^{\tau_b} \frac{\int_s^{\tau_b} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \cdot \left(p_N^{(i)}(s) - p^{(i)} \right) dS^{(i)}(s) \right. \quad \text{(C.9)}$$

$$\left. + \int_{\tau_b}^{\tau} \frac{p^{(i)} \cdot \int_s^{\tau} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \left(dS_X^{(i)}(s) - dS^{(i)}(s) \right) \right) \quad \text{(C.10)}$$

Proof of Theorem B.
As earlier, we will use $U_{b,n}$, $U_{s,n}(\hat{Q})$, and $p^{(i)}$ instead of $U_{b,n}(\tau_b)$, $U_{s,n}(\tau_0, \tau; \hat{Q})$ and $p^{(i)}(\tau_b)$ for short. Since $U_{s,n}(\hat{Q}) - U_{s,n}(Q) \xrightarrow{p} 0$ as $n \to +\infty$ (see Lemma 3.1 in Appendix C.1), it is sufficient to find the asymptotic covariance of $\mathrm{Cov}(U_{b,n}, U_{s,n}(Q))$.

$$\begin{aligned}
\mathrm{Cov}(U_{b,n}, U_{s,n}(Q)) &= \frac{n^{(0)} n^{(1)}}{n} \mathrm{Cov}\left(\left(\widehat{p}^{(1)} - \widehat{p}^{(0)} \right), \quad \int_{\tau_0}^{\tau} Q(t) \left(\hat{S}^{(1)}(t) - \hat{S}^{(0)}(t) \right) dt \right) \\
&= \frac{n^{(0)} n^{(1)}}{n} \sum_{i=0,1} \mathrm{Cov}\left(\int_{\tau_0}^{\tau} Q(t) \hat{S}^{(i)}(t) dt, \hat{p}^{(i)} \right) \\
&= \frac{n^{(0)} n^{(1)}}{n} \sum_{i=0,1} \mathrm{E}\left((\hat{p}^{(i)} - p^{(i)}) \cdot \left(\int_{\tau_0}^{\tau} Q(t) (\hat{S}^{(i)}(t) - S^{(i)}(t)) dt \right) \right)
\end{aligned} \quad \text{(C.11)}$$

To derive (C.11), we use the following results:

- The approximation

$$\sqrt{n^{(i)}} \left(\hat{S}^{(i)}(t) - S^{(i)}(t) \right) = -\sqrt{n^{(i)}} S^{(i)}(t) \int_{\tau_0}^{t} \frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)} + o_p(1) \quad \text{for} \quad t \in [\tau_0, \tau] \quad \text{(C.12)}$$

  given by Fleming & Harrington (see Page 98 in(Fleming and Harrington, 1991));
- $\mathrm{E}\left(\int_{\tau_0}^{\tau} Q(t) \hat{S}^{(i)}(t) dt \right) = \int_{\tau_0}^{\tau} Q(t) S^{(i)}(t) dt$ (Gill (1980), pag 38), since $y^{(i)}(\tau) > 0$;
- $\mathrm{E}(\hat{p}^{(i)}) = p^{(i)}$ and denoting by $K^{(i)}(s) = \int_s^{\tau} Q(t) S^{(i)}(t) dt$,

hence we have:

$$\begin{aligned}(\text{C.11}) &= \frac{n^{(0)}n^{(1)}}{n}\sum_{i=0,1}\mathrm{E}\left((\hat{p}^{(i)}-p^{(i)})\cdot\left(-\int_{\tau_0}^{\tau}K^{(i)}(s)\frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)}\right)\right)+o_p(1/\sqrt{n^{(i)}})\\ &= -\frac{n^{(0)}n^{(1)}}{n}\sum_{i=0,1}\mathrm{E}\left(\hat{p}^{(i)}\cdot\left(\int_{\tau_0}^{\tau}K^{(i)}(s)\frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)}\right)\right)+o_p(1/\sqrt{n^{(i)}})\end{aligned}$$

Furthermore:

$$n^{(i)}\mathrm{E}\left(\hat{p}^{(i)}\cdot\left(\int_{\tau_0}^{\tau}K^{(i)}(s)\frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)}\right)\right)=\int_{\tau_0}^{\tau}K^{(i)}(s)\mathrm{E}\left(\hat{p}^{(i)}\frac{d\bar{M}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}\right) \tag{C.13}$$

because:

- $\frac{K^{(i)}(s)}{\bar{Y}^{(i)}(s)}$ is a bounded $\mathcal{F}_s$-predictable process, then $\int_{\tau_0}^{\tau}K^{(i)}(s)\frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)}$ is a martingale and
  $\mathrm{E}\left(\int_{\tau_0}^{\tau}K^{(i)}(s)\frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)}\right)=0$ (Theorem 2.4.4. in Fleming and Harrington (1991));
- the integrand of (C.13) is a measurable function and by Fubini's Theorem, we can interchange the order of the integral.

We now work with the integrand in (C.13):

$$\begin{aligned}\mathrm{E}\left(\hat{p}^{(i)}\frac{d\bar{M}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}\right) &= \mathrm{E}\left(\hat{p}^{(i)}\frac{d\bar{N}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}\right)-n^{(i)}p^{(i)}\lambda^{(i)}(s)ds\\ &= \frac{\mathrm{E}(\hat{p}^{(i)}d\bar{N}^{(i)}(s))}{y^{(i)}(s)}+o_p(1)\end{aligned} \tag{C.14}$$

because

- $d\bar{M}^{(i)}(s)=d\bar{N}^{(i)}(s)-\bar{Y}^{(i)}(s)\lambda^{(i)}(s)ds$;
- using Lemma C.1:

$$\hat{p}^{(i)}\frac{d\bar{N}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}=\hat{p}^{(i)}d\bar{N}^{(i)}(s)\left(\frac{1}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}-\frac{1}{y^{(i)}(s)}+\frac{1}{y^{(i)}(s)}\right)=\frac{\hat{p}^{(i)}d\bar{N}^{(i)}(s)}{y^{(i)}(s)}+o_p(1)$$

We now work out $\mathrm{E}(\hat{p}^{(i)}d\bar{N}^{(i)}(s))$ and use that $X_{ij}dN_{ij}(s)$ are independent and identically distributed and X_{ij} is independent of $dN_{ik}(s)$ $(k\neq j)$:

$$\begin{aligned}
\mathrm{E}(\hat{p}^{(i)} d\bar{N}^{(i)}(s)) &= \mathrm{E}\left(\left(\frac{\sum_{j=1}^{n^{(i)}} X_{ij}}{n^{(i)}}\right)\left(\sum_{k=1}^{n^{(i)}} dN_{ik}(s)\right)\right) \\
&= \frac{1}{n^{(i)}}\mathrm{E}\left(\sum_{j=1}^{n^{(i)}} X_{ij} dN_{ij}(s)\right) + \frac{1}{n^{(i)}}\mathrm{E}\left(\sum_{j=1}^{n^{(i)}}\sum_{k=1,k\neq j}^{n^{(i)}} X_{ij} dN_{ik}(s)\right) \\
&= \frac{1}{n^{(i)}} n^{(i)} \mathrm{E}\left(X_{i1} dN_{i1}(s)\right) + \frac{1}{n^{(i)}} n^{(i)}(n^{(i)}-1)\mathrm{E}\left(X_{i1} dN_{i2}(s)\right) \\
&= \mathrm{E}\left(X_{i1} dN_{i1}(s)\right) + (n^{(i)}-1) p^{(i)} y^{(i)}(s) \cdot \lambda^{(i)}(s) ds \qquad \text{(C.15)}
\end{aligned}$$

In order to calculate $\mathrm{E}\left(X_{i1} dN_{i1}(s)\right)$, we distinguish whether $s \leq \tau_b$ or $s > \tau_b$ and prove that:

$$\mathrm{E}\left(X_{i1} dN_{i1}(s)\right) = \begin{cases} p_N^{(i)}(s) \cdot y^{(i)}(s) & s \leq \tau_b \\ p^{(i)} J_X^{(i)}(s) \lambda_X^{(i)}(s) ds & s > \tau_b \end{cases} \qquad \text{(C.16)}$$

where $p_N^{(i)}(s) = P\left(X_{i1} = 1 | dN_{i1}(s) = 1\right)$ and $J_X^{(i)}(s) = \mathrm{E}(Y_{i1}(s) | X_{i1} = 1)$.

Indeed, for $s \leq \tau_b$:

$$\mathrm{E}\left(X_{i1} dN_{i1}(s)\right) = P\left(X_{i1} = 1, dN_{i1}(s) = 1\right) = p_N^{(i)}(s) \cdot y^{(i)}(s) \cdot \lambda^{(i)}(s) ds$$

for $s > \tau_b$:

$$\begin{aligned}
\mathrm{E}\left(X_{i1} dN_{i1}(s)\right) &= P(X_{i1} = 1, dN_{i1}(s) = 1) = p^{(i)} \cdot \mathrm{E}\left(dN_{i1}(s) | X_{i1} = 1\right) \\
&= p^{(i)} \cdot \mathrm{E}\left(Y_{i1}(s) | X_{i1} = 1\right) \lambda_X^{(i)}(s) ds = p^{(i)} J_X^{(i)}(s) \lambda_X^{(i)}(s) ds
\end{aligned}$$

By substituting (C.16) in (C.15), we obtain:

$$\mathrm{E}(\hat{p}^{(i)} d\bar{N}^{(i)}(s)) = \begin{cases} (p_N^{(i)}(s) + (n^{(i)}-1) p^{(i)}) y^{(i)}(s) \cdot \lambda^{(i)}(s) ds & s \leq \tau_b \\ p^{(i)} \left(\lambda_X^{(i)}(s) ds \cdot J_X^{(i)}(s) + (n^{(i)}-1) y^{(i)}(s) \cdot \lambda^{(i)}(s) ds\right) & s > \tau_b \end{cases}$$

and then (C.14) becomes:

$$\mathrm{E}\left(\hat{p}^{(i)} \frac{d\bar{M}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}}\right) = \begin{cases} (p_N^{(i)}(s) - p^{(i)}) \lambda^{(i)}(s) ds + o_p(1) & s \leq \tau_b \\ \frac{p^{(i)}}{y^{(i)}(s)} \left(J_X^{(i)}(s) \lambda_X^{(i)}(s) ds - y^{(i)}(s) \lambda^{(i)}(s) ds\right) + o_p(1) & s > \tau_b \end{cases} \qquad \text{(C.17)}$$

Collecting equations (C.11),(C.13),(C.17):

$$
\begin{aligned}
\mathrm{Cov}(U_{b,n}, U_{s,n}(Q)) &= \frac{n^{(0)}n^{(1)}}{n} \sum_{i=0,1} \mathrm{E}\left((\hat{p}^{(i)} - p^{(i)}) \cdot \left(\int_{\tau_0}^{\tau_b} Q(t)(\hat{S}^{(i)}(t) - S^{(i)}(t))dt \right.\right. \\
&\quad \left.\left. + \int_{\tau_b}^{\tau} Q(t)(\hat{S}^{(i)}(t) - S^{(i)}(t))dt \right) \right) \\
&= -\frac{n^{(0)}n^{(1)}}{n} \sum_{i=0,1} \mathrm{E}\left(\hat{p}^{(i)} \cdot \left(\int_{\tau_0}^{\tau_b} K_{\tau_b}^{(i)}(s) \frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)} \right.\right. \\
&\quad \left.\left. + \int_{\tau_b}^{\tau} K_{\tau}^{(i)}(s) \frac{d\bar{M}^{(i)}(s)}{\bar{Y}^{(i)}(s)} \right) \right) + o_p(1/\sqrt{n^{(i)}}) \\
&= -\sum_{i=0,1} \frac{n - n^{(i)}}{n} \left(\int_{\tau_0}^{\tau_b} K_{\tau_b}^{(i)}(s) \mathrm{E}\left(\hat{p}^{(i)} \cdot \frac{d\bar{M}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}} \right) \right. \\
&\quad \left. + \int_{\tau_b}^{\tau} K_{\tau}^{(i)}(s) \mathrm{E}\left(\hat{p}^{(i)} \cdot \frac{d\bar{M}^{(i)}(s)}{\frac{\bar{Y}^{(i)}(s)}{n^{(i)}}} \right) \right) \\
&\quad + o_p(1/\sqrt{n^{(i)}}) \\
&= -\sum_{i=0,1} \frac{n - n^{(i)}}{n} \left(\int_{\tau_0}^{\tau_b} K_{\tau_b}^{(i)}(s) \left((p_N^{(i)}(s) - p^{(i)}) \lambda^{(i)}(s) ds \right) + \right. \\
&\quad \left. + \int_{\tau_b}^{\tau} K_{\tau}^{(i)}(s) \frac{p^{(i)}}{y^{(i)}(s)} \left(J_X^{(i)}(s) \lambda_X^{(i)}(s) ds - y^{(i)}(s) \lambda^{(i)}(s) ds \right) \right) \\
&\quad + o_p(1/\sqrt{n^{(i)}})
\end{aligned}
$$

where $K_{\tau_*}^{(i)}(t) = \int_t^{\tau_*} Q(u) S^{(i)}(u) du$ ($\tau_* = \tau$ or τ_b). Noticing that $\lambda^{(i)}(s)ds = -\frac{dS^{(i)}(s)}{S^{(i)}(s)}$, $\lambda_X^{(i)}(s)ds = -\frac{dS_X^{(i)}(s)}{S_X^{(i)}(s)}$, $y^{(i)}(t) = S^{(i)}(t)G^{(i)}(t)$ and $J_X^{(i)}(s) = S_X^{(i)}(t)G^{(i)}(t)$, we finally obtain:

$$
\begin{aligned}
\mathrm{Cov}(U_{b,n}, U_{s,n}(Q)) &= \sum_{i=0,1} \frac{n - n^{(i)}}{n} \left(\int_{\tau_0}^{\tau_b} \frac{\int_s^{\tau_b} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \cdot \left(p_N^{(i)}(s) - p^{(i)} \right) dS^{(i)}(s) \right. \\
&\quad \left. + \int_{\tau_b}^{\tau} \frac{p^{(i)} \cdot \int_s^{\tau} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \left(dS_X^{(i)}(s) - dS^{(i)}(s) \right) \right) \\
&\quad + o_p(1/\sqrt{n^{(i)}})
\end{aligned}
$$

Since $U_{s,n}(\hat{Q}) - U_{s,n}(Q) \xrightarrow{p} 0$ as $n \to +\infty$, it follows that

$$\lim_{n\to+\infty} \mathrm{Cov}(U_{b,n}, U_{s,n}(\hat{Q})) = \lim_{n\to+\infty} \mathrm{Cov}(U_{b,n}, U_{s,n}(Q)) = \sigma_{bs},$$

and since $\lim_{n\to+\infty} n^{(i)}/n = \pi^{(i)} \in (0,1)$, we finally obtain:

$$\sigma_{bs} = \sum_{i=0,1} \left(1-\pi^{(i)}\right) \left(\int_{\tau_0}^{\tau_b} \frac{\int_s^{\tau_b} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \cdot \left(p_N^{(i)}(s) - p^{(i)} \right) dS^{(i)}(s) \right.$$
$$\left. + \int_{\tau_b}^{\tau} \frac{p^{(i)} \cdot \int_s^{\tau} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \left(dS_X^{(i)}(s) - dS^{(i)}(s) \right) \right)$$

Then the proof is complete. □

C.2.2 Covariance estimation

We will prove at the end of this section that the probability $p_N^{(i)}(t) = P\left(X_{ij} = 1 | dN_{ij}(t) = 1\right)$ can be approximated by the ratio of two hazard functions $p_N^{(i)}(t) = \lambda_{X,T}^{(i)}(t)/\lambda^{(i)}(t)$ where $\lambda_{X,T}^{(i)}(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} P(X = 1, t \le T_1 < t + \Delta t | T_1 > t)$. Using this alternative expression for $p_N^{(i)}(t)$ and noticing that $\lambda^{(i)}(s)ds = -dS^{(i)}(s)/S^{(i)}(s)$, σ_{bs} given in (C.9) can be rewritten as:

$$\sigma_{bs} = \sum_{i=0,1} \left(1-\pi^{(i)}\right) \left(-\int_{\tau_0}^{\tau_b} \left(\int_s^{\tau_b} Q(t) S^{(i)}(t) dt \right) \cdot \lambda_{X,T}^{(i)}(s) ds \right.$$
$$- \int_{\tau_0}^{\tau_b} \frac{\int_s^{\tau_b} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \cdot p^{(i)} dS^{(i)}(s)$$
$$\left. + \int_{\tau_b}^{\tau} \frac{p^{(i)} \cdot \int_s^{\tau} Q(t) S^{(i)}(t) dt}{S^{(i)}(s)} \left(dS_X^{(i)}(s) - dS^{(i)}(s) \right) \right)$$

A consistent estimator for σ_{bs} is obtained by replacing $S^{(i)}(t)$, $S_X^{(i)}(s)$ by the corresponding Kaplan-Meier estimators of $\hat{S}^{(i)}(t)$, $\hat{S}_X^{(i)}(s)$; $Q(t)$ by $\hat{Q}(t)$; and finally replacing $\lambda_{X,T}^{(i)}(s)$ by a kernel function estimator (Andersen et al., 1992).

Derivation of $p_N^{(i)}(t) = \frac{\lambda_{X,T}^{(i)}(t)}{\lambda^{(i)}(t)}$:
We first write:

$$p_N^{(i)}(t) = P\left(X_{ij} = 1 | dN_{ij}(t) = 1\right) = \frac{P\left(X_{ij} = 1, dN_{ij}(t) = 1\right)}{P\left(dN_{ij}(t) = 1\right)}$$

and now:

- $P\left(X_{ij} = 1, dN_{ij}(t) = 1\right) = y^{(i)}(t)\lambda_{X,T}^{(i)}(t)$ because:

  $$P\left(X_{ij} = 1, dN_{ij}(t) = 1\right) = E\left(X_{ij} dN_{ij}(t)\right) = E\left(E\left(X_{ij} dN_{ij}(t) | Y_{ij}(t)\right)\right)$$

  and

  $$\begin{aligned} E\left(X_{ij} dN_{ij}(t) | Y_{ij}(t)\right) &= Y_{ij}(t) P\left(X_{ij} dN_{ij}(t) = 1 | Y_{ij}(t)\right) \\ &= Y_{ij}(t) \frac{P\left(X_{ij} = 1, dN_{ij}(t) = 1, Y_{ij}(t) = 1\right)}{P\left(Y_{ij}(t) = 1\right)} \\ &= Y_{ij}(t) \frac{P(X_{ij} = 1, t \leq T_{ij} < t + dt, T_{ij} > t, C_{ij} > t)}{P(T_{ij} > t, C_{ij} > t)} \\ &= Y_{ij}(t) P(X_{ij} = 1, t \leq T_{ij} < t + dt | T_{ij} > t) \end{aligned}$$

  hence:

  $$\begin{aligned} P\left(X_{ij} = 1, dN_{ij}(t) = 1\right) &= y^{(i)}(t) P(X_{ij} = 1, t \leq T_{ij} < t + dt | T_{ij} > t) \\ &\cong y^{(i)}(t) \lambda_{X,T}^{(i)}(t) dt \end{aligned} \tag{C.18}$$

  because: $E\big(X_{ij} \cdot (N_{ij}(t + \Delta t) - N_{ij}(t)) | Y_{ij}(t) = 1\big) = \lambda_{X,T}^{(i)}(t)\Delta t + o(\Delta t)$.
- $P\left(dN_{ij}(t) = 1\right) = y^{(i)}(t)\lambda^{(i)}(t)dt$ because:

  $$P\left(dN_{ij}(t) = 1\right) = P\left(Y_{ij}(t) = 1\right) E\left(N(t + dt) - N(t) | Y_{ij}(t)\right) \cong y^{(i)}(t)\lambda^{(i)}(t)dt \tag{C.19}$$

  because $E(N_{ij}(t + \Delta t) - N_{ij}(t) | Y_{ij}(t) = 1) = \lambda^{(i)}(t)\Delta t + o(\Delta t)$ (Andersen et al. (1992)).

Taking (C.18) and (C.19) into account, we finally obtain:

$$p_N^{(i)}(t) = \frac{P(X_{ij} = 1, t \leq T_{ij} < t + dt | T_{ij} > t)}{P\left(dN_{ij}(t) = 1\right)} \cong \frac{\lambda_{X,T}^{(i)}(t)}{\lambda^{(i)}(t)}$$

C.3 Additional results case study

In the illustration section, we used the pooled variance estimator for the $\mathcal{L}$-statistic. Next, we show the results by using the pooled and bootstrap variance estimator.

We employ the function `lstats` to compute the $\mathcal{L}$-statistic using the unpooled variance estimator as follows:

```
lstats(time=data$time, status=data$status,
binary=data$binary, treat=data$treat,
tau0=0, tau=4, taub=0.5, rho=0, gam=1, eta=1,
wb=0.25, ws=0.75,
var_est = "Unpooled")

##
## $LTest
## Parameter                Value
## 1 (Standardized) L-Test 4.8550922
## 2                L-Test 3.8415352
## 3    Standard deviation 0.7912384
##
## $Binary_Tests
## Parameter                 Value
##      Standardized L-Test 2.3073454
## Ub           Binary Test 0.4540763
## sd    Standard deviation 0.2431064
##
## $Survival_Tests
## Parameter                Value
##       Standardized Test 4.3529318
## Us         Survival Test 2.4398019
## sd   Standard deviation 0.5604962
##
## $Covariance
## Parameter    Value
## 1 Covariance 0.0003844938
```

We use the function `lstats_boots` to calculate the standardized $\mathcal{L}$-statistic using the bootstrap variance estimator:

```
lstats_boots(data$time, data$status, data$binary, data$
   treat,
tau=4, rho=0, gam=1, eta=1,
wb=0.25, ws=0.75, Boot = 100)

##
## $LTest
## Parameter     Value
```

```
## 1 (Standardized) L-Test 3.9572756
## 2                L-Test 3.8415352
## 3    Standard deviation 0.9707525
##
## $Binary_Tests
## Parameter     Value
## Test Standardized L-Test 2.3073454
## Ub           Binary Test 0.4540763
## sd    Standard deviation 0.1967960
##
## $Survival_Tests
## Parameter     Value
## Test  Standardized Test 4.3529318
## Us         Survival Test 2.4398019
## sd   Standard deviation 0.5604962
```

C.4 Additional results simulation study

As mentioned in the simulation section, we have also evaluated what powers we would have had if instead of having an endpoint with no effect (cases 2 and 3), we have had an endpoint with a small effect. To do so, we considered the parameter configuration that we used in case 2 but instead of considering equal survival curves, we considered $HR = 0.85$; and the parameter configuration in case 3 but instead of $d = 0$ we used $d = 0.05$.

The next two figures show the boxplots of the empirical power when the binary endpoint has small treatment effect (Figure C.1) and when the survival endpoint has small treatment effect (Figure C.2). We observe that the powers when the endpoints are equally important, the $\mathcal{L}$-statistics are still more powerful than the Bonferroni procedure.

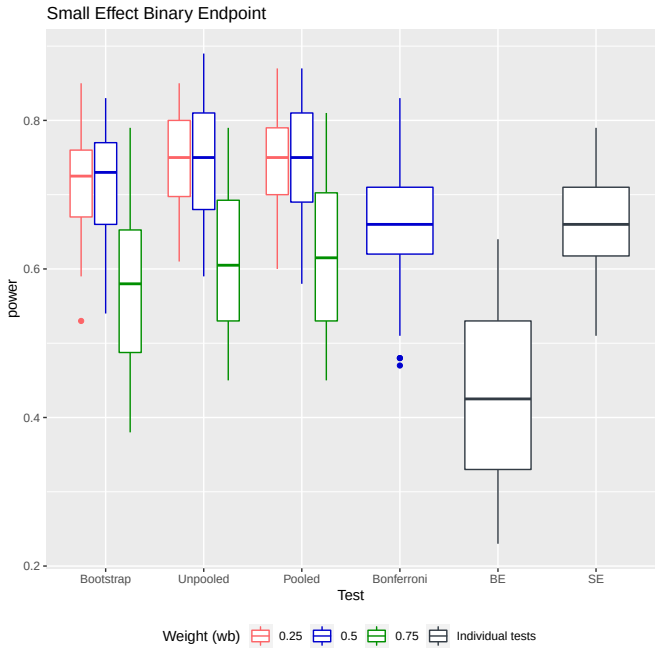


Fig. C.1 Boxplots empirical powers when the binary endpoint has small treatment effect.

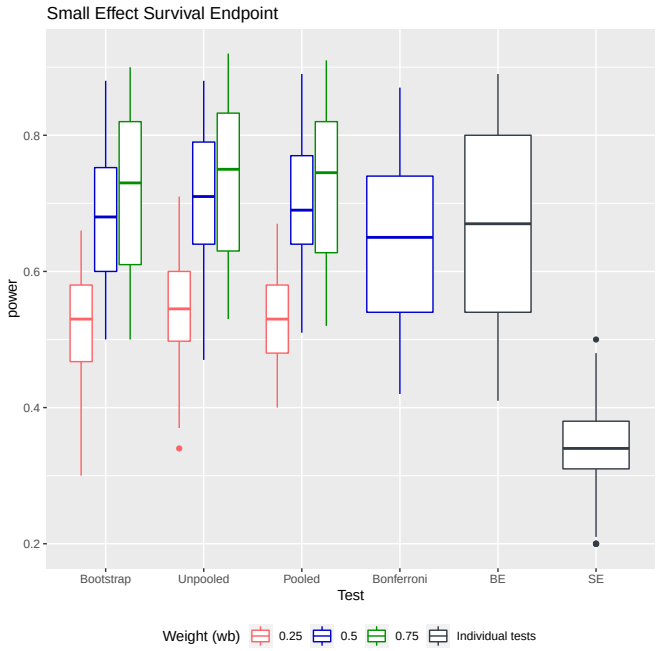


Fig. C.2 Boxplots empirical powers when the survival endpoint has small treatment effect.

List of Figures

List of Tables

Bibliography

Akacha, M., Bretz, F., Ruberg, S. (2017). Estimands in clinical trials – broadening the perspective. *Statistics in Medicine*, **36(1)**, 5-19.

Alosh, M., Bretz, F., and Huque, M. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, **33(4)**, 693–713.

Ananthakrishnan, R., and Menon, S. (2013). Design of oncology clinical trials: A review. *Critical Reviews in Oncology/Hematology* **88(1)**, 144–153.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer, New York.

Anderson, H. V., Cannon, C. P., Stone, P. H., Williams, D. O., McCabe, C. H., Knatterud, G. L., Thompson, B., Willerson, J. T., Braunwald, E., for the TIMI-IIIB Investigators. One-year results of the Thrombolysis in Myocardial Infarction (TIMI) IIIB clinical trial. (1995). A randomized comparison of tissue-type plasminogen activator versus placebo and early invasive versus early conservative strategies in unstable angina and non-Q-wave myocardial infarction. *Journal of the American College of Cardiology*, **26**, 1643– 1650.

Ando, Y., Hamasaki, T., Evans, S.R., Asakura, K., Sugimoto, T., Sozu, T., Ohno, Y. (2015). Sample Size Considerations in Clinical Trials when Comparing Two Interventions using Multiple Co-Primary Binary Relative Risk Contrasts. *Statistics in Biopharmaceutical Research*, **7(2)**, 81–94.

Andrés, A. M., Tejedor, H. (2009). Comments on "How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial". *Statistics in Medicine*, **28**, 173–179.

Asakura, K., Hamasaki, T., Evans, S. R. (2017). Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. *Biometrical Journal*, **59(4)**, 703–731.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *Stanford University Press*. 158–168.

Bauer K. A., Eriksson B. I., Lassen M. R., Turpie A. G.; Steering Committee of the Pentasaccharide in Major Knee Surgery Study. (2001). Fondaparinux compared with

Enoxaparin for the prevention of venous thromboembolism after elective major knee surgery. *New England Journal of Medicine*, **345**, 1305–1310.

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine*. **10**, 871–890.

Boden, W. E., O'Rourke, R. A., Crawford, M. H., Blaustein, A. S., Deedwania, P. C., Zoble, R. G., Wexler, L. F., Pepine, C. J., Ferry, D. R., Chow, B. K., Lavori, P. W., for the Veterans Affairs Non-Q-Wave Infaction Strategies in Hospital (VANQWISH) Trial Investigators. (1998) Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative strategy. *New England Journal of Medicine*, **338**, 1785–1792.

Bofill Roig, M., Gómez Melis, G. (2020). A class of two-sample nonparametric statistics for binary and time-to-event outcomes. arXiv:2002.01369 [stat.ME]. `https://arxiv.org/abs/2002.01369`.

Bofill Roig, M., Cortés Martínez, J., Gómez Melis, G. (2020). Decision tool and Sample Size Calculator for Composite Endpoints. arXiv:2001.03396 [stat.AP]. `https://arxiv.org/abs/2001.03396`.

Bofill Roig, M., Gómez Melis, G. (2019). A new approach for sizing trials with composite binary endpoints using anticipated marginal values and accounting for the correlation between components. *Statistics in Medicine*, **38(11)**, 1935–1956.

Bofill Roig, M., Gómez Melis, G. (2018). Selection of composite binary endpoints in clinical trials. *Biometrical Journal*, **60(2)**, 246–261.

Buoen, C., Bjerrum, O. J., Thomsen, M. S. (2005). How first-time-in-human studies are being performed: a survey of phase 1 dose-escalation trials in healthy volunteers published between 1995 and 2004. *The Journal of Clinical Pharmacology*. **45**, 1123–1136.

Burzykowski, T., Buyse, M. (2006). Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, **5(3)**, 173–186.

Buyse, M., Burzykowski, T., Saad, E. D. (2018). The search for surrogate endpoints for immunotherapy trials. *Annals of Translational Medicine*, **6(11)**, 231–231.

Buyse, M., Piedbois, P. (1996). On the relationship between response to treatment and survival time. *Statistics in Medicine*, **15(24)**, 2797–2812.

Buyse, M., Ryan, L. M. (1987). Issues of efficiency in combining proportions of deaths from several clinical trials. *Statistics in Medicine*, **6(5)**, 565–576.

Bland, J. M., and Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *The BMJ*. **Jan 21**; 310(6973):170.

Bittman, R. M., Romano, J. P., Vallarino, C., and Wolf, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, **96(2)**, 399–410.

Broglio, K. R., Quintana, M., Foster, M., Olinger, M., McGlothlin, A., Berry, S. M., Boileau, J. F., Brezden-Masley, C., Chia, S., Dent, S., Gelmon, K., Paterson, A., Rayson, D., Berry, D. A. (2016). Association of pathologic complete response to

neoadjuvant therapy in HER2-positive breast cancer with long-term outcomes ameta-analysis. *Journal of the American Medical Association Oncology*, **2(6)**, 751-760.

Cannon, C. P., Weintraub, W. S., Demopoulos, L. A., Vicari, R., Frey, M. J., Lakkis, N., TACTICS (Treat Angina with Aggrastat and Determine Cost of Therapy with an Invasive or Conservative Strategy)–Thrombolysis in Myocardial Infarction 18 Investigators. (2001). Comparison of Early Invasive and Conservative Strategies in Patients with Unstable Coronary Syndromes Treated with the Glycoprotein IIb/IIIa Inhibitor Tirofiban. *New England Journal of Medicine*, **344(25)**, 1879–1887.

Cannon, C. P., Weintraub, W. S., Demopoulos, L. A., Robertson, D. H., Gormley, G. J., Braunwald, E. (1998). Invasive versus conservative strategies in unstable angina and non-Q-wave myocardial infarction following treatment with tirofiban: rationale and study design of the international TACTICS-TIMI 18 Trial. *The American Journal of Cardiology*, 82(6), 731–6.

Chen, B. E., and Wang, J. (2020). Joint modeling of binary response and survival for clustered data in clinical trials. *Statistics in Medicine*, **39(3)**, 326–339.

Chow, S. C., Shao, J., Wang, H. (2008). *Sample size calculations in clinical research*. 2nd Edition. Chapman & Hall/CRC Press, Boca Raton, Florida.

Collins, R., Bowman, L., Landray, M., Peto, R., Nonrandomized. (2020). The Magic of Randomization versus the Myth of Real-World Evidence. *New England Journal of Medicine*, **Feb 13; 382(7)**, 674–678.

Cook, T. D., DeMets, D. L. (2008). *Introduction to Statistical Methods for Clinical Trials*. 1st Edition. Chapman & Hall/CRC Press, Boca Raton, Florida.

Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., Bonnefoi, H., Cameron, D., Gianni, L., Valagussa, P., Swain, S. M., Prowell, T., Loibl, S., Wickerham, D. L., Bogaerts, J., Baselga, J., Perou, C., Blumenthal, G., Blohmer, J., Von Minckwitz, G. (2014). Pathological complete response and long-term clinical benefit in breast cancer: The CTNeoBC pooled analysis. *The Lancet*, **384(9938)**, 164–172.

Crans, G. D., Shuster, J. J. (2008). How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. *Statistics in Medicine*, **27**, 3598-3611.

Cuzick, J. (1982). The Efficiency of the Proportions Test and the Logrank Test for Censored Survival Data. *Biometrics*, **38(4)**, 1033–1039.

de Jong, W., Aerts, J., Allard, S., Brander, C., Buyze, J., Florence, E., van Gorp, E., Vanham, G., Leal, L., Mothe, B., Thielemans, K., Plana, M., Garcia, F., Gruters, R., and iHIVARNA consortium. (2019). iHIVARNA phase IIa, a randomized, placebo-controlled, double-blinded trial to evaluate the safety and immunogenicity of iHIVARNA-01 in chronically HIV-infected patients under stable combined antiretroviral therapy. *Trials*. **20**(1):361.

De Martini, D. (2019). Empowering phase II clinical trials to reduce phase III failures. *Pharmaceutical Statistics*, **November 2017**, 1–9.

De Michele, A., Yee, D., Berry, D. A., Albain, K. S., Benz, C. C., Boughey, J., Buxton, M., Chia, S. K., Chien, A. J., Chui, S. Y., Clark, A., Edmiston, K., Elias, A. D., Forero-Torres, A., Haddad, T. C., Haley, B., Haluska, P., Hylton, N. M., Isaacs, C., Esserman, L. J. (2015). The neoadjuvant model is still the future for drug development in breast cancer. *Clinical Cancer Research*, **21(13)**, 2911-2915.

Dmitrienko, A., and Agostino, R. D. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, **32(29)**, 5172–5218.

Donner, A. (1984) Approaches to sample size estimation in the design of clinical trials-a review. *Statistics in Medicine*, **3(3)**, 199–214.

Elashoff, R. M., Li, G., Zhou, Y. (2012). Nonparametric inference for assessing treatment efficacy in randomized clinical trials with a time-to-event outcome and all-or-none compliance. *Biometrika*, **99(2)**, 393-404.

European Medicines Agency. Guideline on multiplicity issues in clinical trials. June 2017. `https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf`

European Medicines Agengy Commitee For Proprietary Medicinal Products (CPMP). Guideline on multiplicity issues in clinical trials. 2016.

Fagerland, M. W., Lydersen, S., Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, **24(2)**, 224–254.

Ferreira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D. M., Montori, V. M., Alonso-Coello, P., Walter, S. P., Guyatt, G. H. (2007). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*, **60(7)**, 651–657.

Ferreira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D. M., Montori, V. M., Alonso-Coello, P., Walter, S. P., Guyatt, G. H. (2007). Composite endpoints in clinical trials: the trees and the forest. *Journal of Clinical Epidemiology*, **60(7)**, 660–661.

Fleiss, J.L. Statistical Methods for Rates and Proportions, Wiley, New York, 1981.

Fleming, T. R., and Harrington, D. P. (1991). Counting Processes and Survival Analysis, volume 8. Wiley Online Library.

Food and Drug Agency. Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics. Guidance for Industry. September 2018.

Food and Drug Administration. Multiple Endpoints in Clinical Trials Guidance for Industry. January 2017.

Food and Drug Administration. Guidance for Industry Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval. October 2014.

Friedman, L. M., Furberg, C. D., DeMets, D. L. (2010). *Fundamentals of Clinical Trials*. 4th Edition. Springer, New York.

Gill, R. D. (1980). Censoring and Stochastic Integrals, Tract 124. Amsterdam: The Mathematical Centre.

Gómez-Mateu, M., Gómez Melis, G. (2017). Clinical trial designs using CompARE. An on-line exploratory tool for investigators. Technical report (UPC). `http://hdl.handle.net/2117/104928`

Gómez, G. and Gómez-Mateu, M. (2014). The asymptotic relative efficiency and the ratio of sample sizes when testing two different null hypotheses. *SORT*, **38(1)**, 73–88.

Gómez, G., Lagakos, S. W. (2013). Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in Medicine*, 32(5):719–738.

Gómez Melis, G. (1986). *Estimation of the time-to-tumor distribution in serial/sacrifice experiments*. Ph.D Thesis, Columbia University.

Gu, M., Follmann, D., and Geller, N. L. (1999). Monitoring a general class of two-sample survival statistics with applications. *Biometrika*, **86(1)**, 45–57.

Hernán, M.A., Robins, J.M. (2020). Causal Inference: What If. Chapman & Hall/CRC Press, Boca Raton, Florida.

Hess, K., and Gentleman, R. (2019). R Package 'muhaz': Hazard Function Estimation in Survival Analysis. Version 1.2.6.1.

Hodi, F.S., O'Day, S.J., McDermott, D.F., Weber, R.W., Sosman, J.A., Haanen, J.B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J.C., Akerley, W., van den Eertwegh, A.J., Lutzky, J., Lorigan, P., Vaubel, J.M., Linette, G.P., Hogg, D., Ottensmeier, C.H., Lebbé, C., Peschel C., Quirt, I., Clark, J.I., Wolchok, J.D., Weber, J.S., Tian, J., Yellin, M.J., Nichol, G.M., Hoos, A., and Urba, W.J. (2010). Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *The New England journal of medicine*. **363(8)**, 711–723.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, **50(3)**, 346–363.

International Council for Harmonization (ICH). ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. International Conference on Harmonisation.

International Council for Harmonization (ICH). Guideline: Statistical principles for clinical trials (E9). 1999.

Karrison, T. G., Ratain, M. J., Stadler, W. M., Rosner, G. L. (2012). Estimation of progression-free survival for all treated patients in the randomized discontinuation trial design. *American Statistician*, **66(3)**, 155–162.

Kieser, M., Kierchner, M., Dölger, E., Götte, H. (2018). Optimal planning of phase II / III programs for clinical trials with multiple endpoints. *Pharmaceutical Statistics*, **January**, 1–21.

Lachin, J. M. (1981). Introduction to Sample Size determination and Power analysis for Clinical Trials. *Controlled Clinical Trials*, **2**, 92–113.

Lai, X., and Zee, B. C. Y. (2015). Mixed response and time-to-event endpoints for multistage single-arm phase II design. *Trials*, **16(1)**, 1–10.

Lai, T. L., Lavori, P. W., and Shih, M. C. (2012). Sequential design of phase II-III cancer trials. *Statistics in Medicine*, **31(18)**, 1944–1960.

Lehmann, E. L., Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd Edition. Springer, New York.

Legler J.M., Lefkopoulou M. and Ryan L.M. (1995). Efficiency and Power of Tests for Multiple Binary Outcomes. *Journal of the American Statistical Association*, **430(90)**, 680–693.

Lefkopoulou M. and Ryan L. (1993). Global tests for multiple binary outcomes. *Biometrics*, **1 49(4)**, 975–88.

Logan, B. R., Klein, J. P., and Zhang, M. J. (2008). Comparing treatments in the presence of crossing survival curves: An application to bone marrow transplantation. *Biometrics*, **64(3)**, 733–740.

Luo, X., Tian, H., Mohanty, S., Tsai, W.Y. (2015). An alternative approach to confidence interval estimation for the win ratio statistic. *Biometrics*, **71(1)**, 139–145.

Magnusson, B. P., Schmidli, H., Rouyrre, N., Scharfstein, D. O. (2019). Bayesian inference for a principal stratum estimand to assess the treatment effect in a subgroup characterized by postrandomization event occurrence. *Statistics in Medicine*, **38(23)**, 4761-4771.

Marsal, J. R., Ferreira-González, I., Bertran, S., Ribera, A., Permanyer-Miralda, G., García-Dorado, D., Gómez, G. (2017). The Use of a Binary Composite Endpoint and Sample Size Requirement: Influence of Endpoints Overlap. *American Journal of Epidemiology*, **185(9)**, 832–841.

Mascha E.J. and Sessler D.I. (2011). Statistical grand rounds: design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research. *Anesthesia & Analgesia*, **112(6)**, 1461–71.

Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis. *The BMJ*, **2**, 769–782.

Mick, R., and Chen, T.T. (2015). Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies. *Cancer Immunology Research*, **3(12)**, 1292–1298.

Muller, H.G., and Wang, J.L. (1994). Hazard Rate Estimation under Random Censoring with Varying Kernels and Bandwidths. *Biometrics*, **50(1)**, 61–76.

Noether, G. E. (1954). On a Theorem of Pitman. *The Annals of Mathematical Statistics*.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, **04**, 1079–1087.

Papageorgiou, G., Mauff, K., Tomer, A., Rizopoulos, D. (2019). An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annual Review Of Statistics and Its Application*, **6(15)**, 1–18.

Pepe, M. S., and Fleming, T. R. (1989). Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data. *Biometrics*, **45(2)**, 497–507.

Pepe, M. S., and Fleming, T. R. (1991). Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53(2)*, 341–352.

Pipper, C. B., Ritz, C., and Bisgaard, H. (2012). A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, **61(2)**, 315–326.

Pocock, S. J., McMurray, J. J. V., Collier, T.J. (2015). Statistical Controversies in Reporting of Clinical Trials Part 2 of a 4-Part Series on Statistics for Clinical Trials. *Journal of the American College of Cardiology*, **66(23)**, 2648–2662.

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

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, **43(3)**, 487–498.

Pocock, S. J. (1983). Clinical trials: a practical approach. *John Wiley & Sons.*

Prentice RL. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44(4)**, 1033-48.

Rauch, G., Kunzmann, K., Kieser, M., Wegscheider, K., König, J., Eulenburg, C. (2017). A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance. *Statistics in Medicine*, **35(5)**, 749–767.

Rauch, G., Jahn-Eimermacher, A., Brannath, W. and Kieser, M. (2014). Opportunities and challenges of combined effect measures based on prioritized outcomes. *Statistics in Medicine*, **33(7)**, 1104–1120.

Rauch, G., Wirths, M. and Kieser, M. (2014). Consistency-adjusted alpha allocation methods for a time-to-event analysis of composite endpoints. *Computational Statistics and Data Analysis*, **75**, 151–161.

Rauch, G. and Kieser, M. (2013). An expected power approach for the assessment of composite endpoints and their components. *Computational Statistics and Data Analysis*, **60(1)**, 111–122.

Rauch, G., Kieser, M. Multiplicity adjustment for composite binary endpoints. *Methods of Information in Medicine*. 2012;51(4):309–317.

Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data, with Applications in R. Boca Raton: Chapman & Hall/CRC.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure period-application to control of the health worker survivor effect. *Mathematical Modeling*, *7*, 1393–1512 (and Addendum).

Romano, J. P., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, **100(469)**, 94–108.

Rosenblatt, M. (2017). The Large Pharmaceutical Company Perspective. *New England Journal of Medicine*, **376(1)**, 52–60.

Rufibach, K. (2018). Treatment effect quantification for time-to-event endpoints – Estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*, **October 2017**, 1-21.

Ryan, L. M. (1985). Efficiency of Age-Adjusted Tests in Animal Carcinogenicity Experiments. *Biometrics*, **41(2)**, 525–531.

Sander, A., Rauch, G., Kieser, M. (2016). Blinded sample size recalculation in clinical trials with binary composite endpoints. *Journal of Biopharmaceutical Statistics*.

Shen, Y., and Fleming, T. R. (1997). Weighted mean survival test statistics: A class of distance tests for censored survival data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **59(1)**, 269–280.

Shen, Y., and Cai, J. (2001). Maximum of the weighted Kaplan-Meier tests with application to cancer prevention and screening trials. *Biometrics*. **57(3)**, 837–843.

Senn, S. S. (2008). Statistical issues in drug development. 2n Edition. John Wiley & Sons.

Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*. **6**, 161–170.

Sievert, C. (2020) *Interactive web-based data visualization with R, plotly, and shiny*. `https://plotly-r.com/index.html`]

Sozu, T., Sugimoto, T., Hamasaki, T., Evans, S.R. (2015). *Sample Size Determination in Clinical Trials with Multiple Endpoints*. 1st Edition. Springer.

Sozu, T., Sugimoto, T., Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*, **29(21)**, 2169–2179.

Storer, B. (1989). Design and analysis of phase I clinical trials. *Biometrics*, **45**, 925–937.

Stone, G.W., Ellis, S. G., Cannon, L., Mann, J. T., Greenberg, J.D., Spriggs D, O'Shaughnessy, C.D., DeMaio, S., Hall, P., Popma, J.J., Koglin, J., Russell, M.E.; TAXUS V Investigators. (2005). Comparison of a polymer-based paclitaxel-eluting stent with a bare metal stent in patients with complex coronary artery disease: a randomized controlled trial. *Journal of the American Medical Association*, **294(10)**, 1215–23.

Sugimoto, T., Hamasaki, T., Evans, S. R., Sozu, T. (2017). Sizing clinical trials when comparing bivariate time-to-event outcomes. *Statistics in Medicine*, **36(9)**, 1363–1382.

Sun, R. (2020). R package. GitHub Repository: `https://github.com/ryanrsun/reconstructkm`.

Thall, P. F., Nguyen, H. Q., Wang, X., Wolff, J. E. (2012). A hybrid geometric phase II/III clinical trial design based on treatment failure time and toxicity. *Journal of Statistical Planning and Inference*, **142(4)**, 944–955.

Thall, P. F. (2008). A review of phase 2-3 clinical trial designs. *Lifetime Data Analysis*, **14(1)**, 37–53.

Tomlinson, G., Detsky, A. S. (2010). Composite End Points in Randomized Trials. *Journal of the American Medical Association*, **303(3)**, 267.

Trivedi, P. K., and Zimmer, D. M. (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, **1(1)**, 1–111.

Tsiatis, A. A., Davidian, M., Holloway, S. T., Laber, E. B. (2020). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. 1st Edition. Chapman & Hall/CRC Press, Boca Raton, Florida.

Tsiatis, A. A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14(3)**, 809–834.

Vardi, Y., Ying, Z. L., Zhang, C. H. (2001). Two-sample tests for growth curves under dependent right censoring. *Biometrika*, **88(4)**, 949–960.

Weedon-Fekjaer, H., Lindqvist, B. H., Vatten, L. J., Aalen, O. O., Tretli, S. (2008). Breast cancer tumor growth estimated through mammography screening data. *Breast Cancer Research*, **10(3)**, 1–13.

Whitt, W. (1976). Bivariate distributions with given marginals. *Annals of Statistics*, **4(6)**, 1280–1289.

Wickham, H., Grolemund, G. (2020) *R for Data Science*. `https://r4ds.had.co.nz/`

Wickham, H. (2020) *Mastering Shiny*. `https://mastering-shiny.org/`

Wickham, H. (2020) *R packages*. `http://r-pkgs.had.co.nz/`

Wiki CMS groupware. `http://info.tiki.org`.

Wilson, M. K., Collyar, D., Chingos, D. T., Friedlander, M., Ho, T. W., Karakasis, K., and Oza, A. M. (2015). Outcomes and endpoints in cancer trials: Bridging the divide. *The Lancet Oncology*, **16(1)**, e43–e52.

Yihui Xie, J. J. Allaire, Garrett Grolemund. (2020) *R Markdown: The Definitive Guide*. `https://bookdown.org/yihui/rmarkdown/`