STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yanning Liu

Aug, 2016

STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF CLINICAL TRIALS

Yanning Liu, Ph. D.

Cornell University 2016

Chapters 1-5 concern statistical methods in designing and analyzing data for survival clinical trials, and predicting trial duration. In Chapter 1, a method is proposed to calculate additional time to event after being censored at the withdrawal time together with some imputation strategies to conduct sensitivity analyses for a real trial with informative censoring. Chapter 2 extends Mehta and Pocock (2010) to provide a method for deciding sample size increase in adaptive survival trials. Chapter 3 is inspired by the need from a real trial. A novel method for predicting the timing of events in clinical trials with survival endpoints is proposed using different parametric event distributions in the presence and absence of censoring. Chapter 4 investigates scenarios in planning a comparative group sequential survival clinical trial with subjects who remain event-free can stay until the study is terminated; Chapter 5 treats the same issues as in Chapter 4 but for survival trials with subjects who have a fixed follow-up time after randomization.

Chapters 6-8 concern statistical methods in clinical trials with sequential parallel designs, which have been proposed for trials with high placebo response rates which can lead to a higher failure rate in drug development. Chapter 6 introduces the extended sequential parallel design (ESPD), in which there is re-randomization of not only placebo non-responders during Period 1 but also of drug responders during

Period 1 into Period 2. Chen et al. [Contemp. Clin. Trials, 32: 592-604 (2011)] heuristically proved that the covariance of two estimators is zero assuming equal correlation coefficients. In Chapter 7, this covariance is re-derived without any strong assumption in equality between two correlation coefficients. Assuming the number of subjects continuing into Period 2 is a random variable, the covariance is re-confirmed to be zero for both normal and binomial data. Chapter 8 clarifies a misunderstanding of a new approach to drug-placebo difference calculation in short term antidepressant-drug trials, which was proposed by Rihmer at al. (2011).

Chapter 9 proposes optimized asymmetric group sequential designs that consider constraints on stopping probabilities at stage one (under the null and alternative hypotheses) in addition to traditional overall type I error and power. Thus validity of a group sequential design is ensured from the very first stage throughout.

Utilizing Box and Muller (1958), one of the most popular methods of generating standard normal random variable using two independent uniform (0, 1) deviates, a new method is proposed in Chapter 10 to combine two p-values from two disjoint samples for designing a trial with two stages.

BIOGRAPHICAL SKETCH

Yanning Liu is currently a principal statistician working at Janssen Pharmaceuticals, Inc. in charge of several phase 2-3 trials for an “accelerated-to-value” (i.e., most important ones in the pipeline) investigational compound to treat treatment-resistant depression (TRD) and suicidality in patients with major depressive disorder (MDD). Since Yanning joined Johnson and Johnson (JNJ) in Jan 2006, she has worked on many phase1-3 trials and has participated three successful compounds’ U.S. and world-wide submissions.

While in Cornell, after finishing required core courses for entering graduate study for field of statistics, Yanning transferred from field of microbiology to field of statistics in 2001. During the subsequent four and a half years, Yanning has finished all required courses, exams and teaching assignments; and finished one summer intern in JNJ in 2014 and the other one in Pfizer Inc. in 2015. Prior to Cornell, Yanning has obtained a Bachelor’s degree in Microbiology and a Master degree in Microbiology and Genetics from China Agricultural University.

Outside academics, Yanning is a paper reviewer for International Journal of Biometrics and Biostatistics and was a former Vice President of Cornell Chinese Student Association. Yanning has been participating volunteer work within JNJ and for nearby communities and has been an active participant in giving presentations within and outside JNJ.

To My Parents

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Bruce W. Turnbull for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in designing and analyzing data in clinical trials, all the knowledge in group sequential and adaptive designs and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study. And I could not have imagined joining in the pharmaceutical industry without him.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Martin Wells and Prof. Giles Hooker, for their insightful comments and encouragement, but also for the hard questions which led me to widen my research from various perspectives.

My sincere thanks also go to Janssen Pharmaceuticals Inc. and my colleagues there, who provided me an opportunity to join their team as an intern then granted me a full time position, and who allowed me to use trial data as examples in the manuscripts. Without they precious support it would not be possible to conduct these research projects. In fact, all research work complied here have been inspired from my daily work at Janssen Pharmaceuticals Inc.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD: Antidepressant
ADRS: Adaptive Dose-Ranging Studies
A-GSD: Adaptive Group Sequential Design
ASN: average sample number
BK: Bauer and Kohne
BLAs: Biologic License Applications
BM: Box and Muller
CDF: Cumulative Distribution Function
CIBIS Cardiac Insufficiency Bisoprolol Study
CRM: Continual Reassessment Method
CROs: Contracted Research Organizations
DB: Double-blind
EDC: Electronic Data Capture
EOS: End-of-study
ESPD: Extended Sequential Parallel Design
FDA: Food and Drug Administration
GSD: Group Sequential Design
HDRS17: Hamilton Depression Rating Scale
HAMD: Hamilton Rating Scale for Depression
IPCW: Inverse Probability-of-censoring Weights
IWRS/IVRS : Interactive Web Response System
KD: Kim-Demets
KM: Kaplan-Meier
MCMC: Markov Chain Monte Carlo
MDD: Major Depressive Disorder
MMRM: Mixed effect Model Repeat Measurement
NDA/BLA: New Drug Application/Biologic License Application
NMEs: New Molecular Entities
OBF: O’Brien and Fleming
PhRMA: the Pharmaceutical Research and Manufacturers of America
PL: Placebo
SPD: Sequential Parallel Design
SSRI: Selective Serotonin Reuptake Inhibitors
TPM: Topiramate
Tufts CSDD: Tufts Center for the Study of Drug Development
WT: Wang and Tsiatis

LIST OF SYMBOLS

Symbols are defined differently in each Chapter

PREFACE

Chapters 1 and 8 have been published at *Journal of Biopharmaceutical Statistics* online (March, 2016) and *Open Journal of Statistics* (2015) respectively with Y. Liu as the sole author. Chapter 2 was published online at July 17th by *Communication in Statistics: Theory and Methods* in January 2016 with Y. Liu as the first author (coauthored with Pilar Lim). Chapter 5 and 7 were accepted by *Communication in Statistics: Theory and Methods* in January 2016 and in Jun 2016, respectively, with Y. Liu as the sole author. Chapters 3, 4, 6 are under review by *Journal of Biopharmaceutical Statistics*, *Statistics in Biopharmaceutical Research and Communication in Statistics: Theory and Methods*, respectively. Chapters 9 and 10 will be submitted shortly. All chapters in this dissertation are original work with Y. Liu as the only author or the first author and thus are eligible to be included as respective chapters in a PhD dissertation per Cornell Graduate School. All published papers or manuscripts accepted or being reviewed or to be submitted have gone through the proper processes for data use and external publication process at Janssen Pharmaceuticals Inc., because Y. Liu is a current employee at Janssen Pharmaceuticals Inc.

CHAPTER 0

Overview of the Dissertation

Section 0.1: Phase 2 and 3 Clinical Trials in Drug Development

The pharmaceutical history can be roughly viewed as consisting of three periods (i.e., mid-1800 to 1945, 1945-1970 and 1970-1980s). Between mid-1800 and 1945, botanicals such as morphine and quinine were extracted; epinephrine, norepinephrine were synthesized for treating asthma attacks as well as nasal congestion and amphetamine synthesized for psychiatric indications; barbiturates were discovered and developed by Bayer pharmaceuticals for treating attention deficit disorder and epilepsy; discovery and widespread availability of insulin therapy has changed the prognosis for diabetics from only having a few months of life expectancy to just being a chronic disease (Rosenfeld L, 2002); anti-infective researches resulted in many classes of antibiotics (for example, Salvarsan, Prontosil and Penicillin) and vaccines so that human beings for the first time in history had a way to substantially reduce death rate after being disastrously infected by bacteria or viruses. In the post-war years, 1945-1970, there were further advancements in anti-infective research and development of antihypertensive drug followed with invention of oral contraceptives, the thalidomide issue and the Kefauver-Harris Amendments. In the years of 1970-1980s, the discovery and development of statins helped the patients reduce cholesterol levels so that their chances of dying of a heart attack would be reduced by 40%. Since 1990, drug discovery and development has entered a new era, focusing on understanding the metabolic pathways related to a disease state or pathogen and finding a molecule interfering these pathways. Now large pharmaceutical corporations participate in the complete range of drug discovery, formulation, development, manufacturing, quality control, marketing, sales and distribution while smaller organizations focus on a smaller spectrum of the whole process such as discovery drug candidates or formulation or clinical development. Drug development consists of the following phases: 1) Preclinical phase to conduct in vitro and in vivo studies in non-human subjects for gathering efficacy, toxicity and pharmacokinetic information; 2) Phase 0 to test on approximately 10 human volunteers to gather pharmacodynamics and pharmacokinetics information; 3) Phase 1 to test the drug on 20-100 healthy volunteers for checking dose range; 4) Phase 2 (on 100-300 patients) to determine whether drug candidate can have any efficacy; 5) Phase 3 (on 1000-2000 patients) to test and confirm drug's therapeutic effect, effectiveness and safety; and 6) Phase 4 for post marketing surveillance and watching drug use in public. In the past decade, the author of this dissertation has been working on trials from phase 1 to phase 4 but focusing on phase 3 trials for registration submission to the Food and Drug Administration (FDA) and other regulatory agencies from the rest of the world. According to PhRMA's homepage (http://www.phrma.org/about), America's biopharmaceutical industry had more than 550 new medicines approved by FDA, which performs the lead role in the world. However, among all investigated compounds for use in humans, only a very small fraction are eventually approved by FDA in the U.S. or other regulatory agencies outside U.S. Accordingly to FDA's website (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/default.ht m), the average number of submitted and approved New Molecular Entities (NMEs) or Biologic License Applications (BLAs) in the U.S. between 2004 and 2013 were 38 and 29 per year with average approval rate of 83.9% in this decade. The inspiring news is that submitted NMEs/BLAs (41 and 45 in calendar years of 2014 and 2015) were all approved with medicines resulting from new advancements in science and technology in the past a couple of decades. On the other hand, the approval comes from substantial investment in pre-human and clinical trials and post-approval safety monitoring. According to the Tufts Center for the Study of Drug Development (Tufts CSDD located at http://csdd.tufts.edu/Research/Milestones.asp) and J.A. DiMasi et al. 2016, the predicted overall clinical success rate is only 11.83%, the majority of the drug candidates will fail during the development process and will then generate no revenue in the end. Hence once the cost of failed drugs are taken into account, the average out-of-pocket cost (not including marketing cost) and capitalized cost (adjusted for the time value of money as well as the cost of debt) are 1,395 and 2,558 million U.S. dollars respectively in 2013 (DiMasi et al. 2016). Among the estimated average total capitalized cost per a NME/BLA in 2013, 1,098 million (43%) was used in the pre-human tests while the rest of 1,460 million (57%) was used for clinical trials (DiMasi et al. 2016). Over the time, the total capitalized cost per a NME/BLA in the decade of interest is always more than twice that of the previous decade. They are 179, 413, 1044 and 2,558 million U.S. dollars in 1970s-early 1980s, 1980s-early 1990s, 1990s–mid 2000s and 2000s –mid 2010s respectively (DiMasi et al. 2016).

Due to the fact that substantial time and cost are needed in developing NMEs/BLAs, innovations and improvements are imperative at every aspect during drug development process. To name a few here, novel and more sophisticated measuring scales; new generation of computers/ workstations with higher computing power; more complicated Electronic Data Capture (EDC) system for data capture and Interactive Web Response System (IWRS/IVRS) for patient enrollment, randomization, medication dispense according to protocol and subject withdrawal; dynamic and real time communication between EDC system and IWRS/IVRS system during trial execution; more multi-site and multi-countries trials; more collaborations among big pharmaceutical organizations, small biotechnology companies and with Contracted Research Organizations (CROs); and innovative statistical methods to address unmet needs in drug development including saving time and cost together with making better use of data information at every step of the drug development. As a clinical biostatistician, the author is more familiar with phase 2 and 3 trials and will briefly discuss some advancement in adaptive designs in Section 0.2 below.

Section 0.2: Adaptive Designs in Clinical Trials

Particular motivation for research and implementation of adaptive designs came from the observation of low transition probability both from phase 2 to phase 3 (36%) and from phase 3 to New Drug Application/Biologic License Application (NDA/BLA) submission (62%) (Fig. 1, J.A. DiMasi et al. 2016), where the low rates could possibly be attributed to reasons such as the inability to demonstrate superiority of an investigational compound over placebo, suboptimal dose selected at phase 2 and incorrect patient population investigated, just to name a few here.

There are four major categories of adaptive designs:

1) Adaptive randomization designs including later randomization based on past

treatment assignment only, or past treatment assignment plus covariate-adaptive, or plus response-adaptive or plus both covariate-adaptive and response-adaptive;

2) Group sequential designs (GSDs). Dating back to the 1920s, sequential design started to assess trial data after every observation, while group sequential designs include a small number of interim analyses as data from groups of subjects become available. By interim results, a trial could stop for efficacy or futility at interim. Design parameters are all specified prior to trial start and are not allowed to be modified during the trial. GSDs have been very popular since 1970 and still popular now;

3) Sample size re-estimation. In contrast to GSDs, sample size re-estimation allows one to adjust the sample size of the trial based on cumulative interim data using either blinded data or un-blinded data. Sample size re-estimation using blinded data is used to update variability of the data for a normal endpoint, or to update response rate in the control group when data are binary or to update baseline hazard rate for the combined group in the trial with survival endpoint. For sample size re-estimation, re-estimated sample size is based on treatment effect calculated using un-blinded interim data, which provides an opportunity to adjust the sample size when the treatment effect was over-estimated a priori;

4) Adaptive dose-response designs occur in phase 1 and 2 trials. This includes continual reassessment method (CRM) to estimate maximum tolerable dose in phase 1 trials. Estimating minimum effect dose using novel methods and simulations are currently under-investigation by the PhRMA “Adaptive Dose-Ranging Studies” (ADRS) working group;

5) Treatment selection designs. Supposing a trial starting with several treatments and a concurrent control, one (or more) dose (doses) are selected based on interim point estimates, results of hypothesis testing, external information and expert knowledge. Selected dose(s) and control groups are continued to stage 2. Data from the two stages will be combined using a combination test to conduct hypothesis testing in a way that the overall type I error is controlled at a pre-specified level, thus providing confirmatory evidence of efficacy to support new drug application or biologic license application. As a clinical biostatistician, the author herself has worked on many phase 2 and 3 trials in the central nervous system (CNS) for a decade and has participated three compounds' U.S. and the rest of the world submissions. In Sections 0.3-0.12, the abstract of ten manuscripts that were triggered by real trial questions will be presented, where Sections 0.4, 0.6, 0.7, 0.11 and 0.12 are about adaptive designs, Section 0.3 and 0.5 are about sensitivity analyses and trial monitoring for survival trials, and Sections 0.8, 0.9 and 0.10 are about a novel design of sequential parallel design to deal with the issue of having high placebo response rate in clinical trials.

Section 0.3: Sensitivity Analyses for Informative Censoring in Survival Data: A Trial Example

In a controlled clinical trial comparing an experimental drug to a control using time to event analysis, the logrank test is normally used to test against the equality between two survival curves when the proportional hazard rate assumption is held, which of course requires non-informative censoring. The authors used an example from a randomized, double-blind, parallel group, low-dose active controlled study comparing the safety and efficacy of two doses (400 mg/day versus 50 mg/day) of study medication used as monotherapy for the treatment of newly diagnosed or recurrent epilepsy. This analysis imputes the event time of subjects considered to have problematic informative censoring to demonstrate the impact of violations in necessary assumptions, and assesses robustness of the p-value as calculated from imputed data as compared with un-imputed data. Assuming a parametric distribution for time to event, had these subjects resulted in an event in the trial after withdrawal, the expected additional time to event is formulated and calculated using methods developed in this paper. Combining the imputed informative censoring subjects with the remainder of the original data, new p-values are obtained using the log-rank test and compared to the original p-value. KM plots are also compared.

Section 0.4: Sample Size Increase during a Survival Trial When Interim Results are Promising

In clinical trials with survival end point, an anticipated log hazard ratio is used to plan a trial (with either fixed sample design or a design with multiple stages) before trial begins. Uncertainty of log hazard ratio under alternative hypothesis may create a need for a sample size increase when interim results are promising and treatment effect has been underestimated. This paper generalizes Mehta and Pocock (2000) method to provide a way for adaptive sample size increase in survival trials. Unlike trials with normal or binary endpoints, subjects who were at risk at the interim analysis contribute both at interim and at final, resulting in dependent data structure between interim log-rank test and final log-rank test. A method to create independent increment in order to obtain a weighted test statistic and search for an adjusted critical value for final analysis is proposed. Before trial start, given the information time for interim analysis and the ratio of maximum total sample size after increase to planned sample size before trial start are specified, the sample space is divided by the observed test statistic at interim into three zones: unfavorable, promising and favorable, the sample size (required number of events) remains unchanged when interim test statistic is located in unfavorable or favorable zones, but is increased if it is located in the promising zone instead. Implementation of sample size increase in survival trials is described in details. Simulations with scenarios with equally spaced group sequential designs with/without censoring and with/without adaptation in sample size are performed. Simulations allowing a 4-fold increase in sample size against 2-fold increase are compared. Besides equally spaced group sequential designs, interims occurring at the earlier part (at 20% of anticipated information is used) or the later part (at 80% of anticipated information is used) are also investigated.

Section 0.5: Prediction of the Timing of Events in Clinical Trials with Survival Endpoints: A Trial Example

In event-based clinical trials, interim and final analyses at pre-specified event times are often proposed. In a randomized withdrawal trial with a time-to-event primary endpoint, the design consists of subjects receiving a test treatment for a specified period and then being randomized to continue on that treatment or placebo. We present methodology to predict the time of reaching a required number of events during the double-blind phase of such a trial. We consider prediction at any time during the course of this trial: at the beginning of the trial; during the open-label phase of the trial and also during the double-blind phase of the trial (where some subjects could still be in the open-label phase). There has been recent work on tackling various aspects of this problem using parametric, semi-parametric or from a Bayesian perspective. Starting from Whitehead's method (2001), we consider four additional features: (i) censoring process can be incorporated; (ii) calculating expected number of events by a future calendar time, t_2, for subjects who were in the risk set at t_1; (iii) predicting number of events by a future time point t_2 for subjects who were enrolled prior to randomization and will be randomized at a fixed time point before t_2; and (iv) various parametric survival distributions other than exponential (i.e., Weibull, Lognormal, Log logistic). We applied our methodology during the conduct of a recently completed clinical trial to accurately predict the timing of the interim analysis. This allowed sufficient resources to be deployed leading to timely data analysis and reporting.

Section 0.6: Planning a Comparative Group Sequential Clinical Trial with Loss to Follow-up and a Period of Continued Observation

This paper is motivated by Rubinstein, et al., (1981) and Kim and Tsiatis (1990) and provides a way to design group sequential trials analyzed using logrank test for comparing survival under two treatments with loss to follow-up and a period of continued observation. These are frequently encountered in Phase II/III clinical trials. A method is developed to calculate the length of accrual period to assure a desired power for given control group median time to event, hazard ratio, length of the period of continued observation, information time of analyses and times of analyses, hazard rate of time to censoring and significance level. The results show that, similar to trials with fixed duration (Rubinstein, et al. 1981), introducing a period of continued observation after the end of patient accrual period reduces the total number of patients required to detect treatment effect substantially. Assuming both time to event and time to censoring (loss to follow-up) are exponential, the estimator of log hazard ratio (placebo vs. treatment) is used to test the null hypothesis of equality in survival distributions between treatment and placebo groups. Tables are created in which total trial durations are calculated for a wide range of cases for O'Brien and Fleming (1979), Pocock (1977) and Wang and Tsiatis(1987) efficacy upper boundaries, respectively. For the same accrual rate, three different curves are depicted to show the impacts of time to censoring and a period of continued observation on accrual time to ensure power in respective group sequential settings.

Section 0.7: Planning the Duration of a Survival Group Sequential Trial with a Fixed Follow-up Time for All Subjects

To account for the need of exploring operating characteristics of survival group sequential trials with a fixed follow-up period for each subject after randomization, the accrual time and total trial duration to ensure power and type I error rate requirements are explained. Situations investigated are for hazard ratios ranging from 1.3 to 3.0, with slow or high accrual rate, and in the presence or absence of censoring. Impacts of hazard rate, accrual rate and competitive censoring on accrual time and subsequently on total trial duration are carefully illustrated by well-designed tables and figures. Real calendar time for interim analyses, needed number of events and recruited number of subjects at time of interim analyses, are also tabulated so that all operation characteristics can be assessed prior to the trial start and re-assessed during the trial after incorporating adjusted accrual rate based on blinded data review. The importance of having such explorations is illustrated via a motivating example.

Section 0.8: Optimal Weighted Z Test and Linear Combination Test in Extended Sequential Parallel Designs

Many times in clinical trials using Sequential Parallel Design (SPD) with two treatments (placebo and drug), subjects are randomized in Period 1 and placebo non-responders are re-randomized in period 2 to either continue with placebo or switch to active drug. The re-randomization of placebo non-responders during Period 1 into Period 2 helps to overcome the potential imbalance in baseline factors resulting due to informative withdrawals during Period 1 was discussed by Chen et al. (2011) and Liu et al. (2012). In this paper, we introduce extended SPD (ESPD) and consider the re-randomization of not only placebo non-responders during Period 1 but also the re-randomization of drug responders during Period 1 into Period 2. Statistical methods to analyze data from an ESPD are discussed. An optimal weighted *Z* test which combines three individual test statistics is suggested to test the hypothesis of no drug effect across periods. It is shown that the ESPD is more efficient compared to SPD. Simulation results are also presented. Additionally, a linear combination test is proposed for binary data, which demonstrates good and fair operational characteristics under both null and alternative hypotheses, respectively.

Section 0.9: Covariance and Variance Evaluations of Two Estimators for Drug-placebo Difference in a Trial with Sequential Parallel Design

Fava et al., 2003 proposed Sequential Parallel Design (SPD) to test for a drug effect in the presence of a placebo effect by combining two estimators from first and second periods of the trial. Here subjects are randomized to receive either placebo or drug in the first period and only placebo non-responders at the end of the first period are continued into the second period. Chen et al. (2011) heuristically proved that the covariance of two estimators is zero assuming the correlation coefficient between the first and the second period normal responses for subjects who were placebo non-responders in period 1 and continued to be treated by placebo in period 2 being the same as the correlation coefficient between the first and the second period normal responses for subjects who were placebo non-responders in period 1 and continued to be treated by testing drug in period 2. However in practice it is often difficult to justify the equality assumption between two correlation coefficients. In this article, the above covariance is re-derived without needing any strong assumption in equality between two correlation coefficients. Assuming number of subjects continuing into period 2 being a random variable, covariance is re-confirmed to be zero not only for normal data but also for binomial data. Subsequently, the sample size for a SPD trial using weighted test for hypothesis testing is derived with estimated non-responder rate at the end of the first period being replaced by its expected value. The efficiency of a SPD design is evaluated accordingly relative to fixed sample design for both scenarios. Simulations are also performed to assess type I error rate and power when period 1 and 2 endpoints are correlated.

Section 0.10: Misunderstanding of a New Approach to Drug-Placebo Difference Calculation in Short Term Antidepressant-Drug Trials

In clinical trials, drug effect is measured by a difference between subjects who are treated by experimental drug against placebo-treated subjects. In case of binary data, with observing YES/NO on each subject in certain period of time, it is the proportion of subjects who respond in treatment group minus the proportion of responders in placebo group (for example, 50% vs. 30%). However, a greater difference was proposed by Rihmer et al. (2011) with their supporting arguments, in that antidepressant response and placebo response had different mechanisms and there were equal chances for antidepressant responder to be responding to placebo and not responding to placebo at all. Therefore, the authors proposed 50% - 30% * 50% when the response rate in the treatment group and the placebo group are 50% and 30% respectively, resulting in higher drug-placebo difference than traditional understanding of 50% - 30%. In this article, we tried to explain why the authors misunderstood the drug-placebo concept for evaluating drug superiority, their misunderstanding of assumptions of traditional calculation, as well as their wrong reasoning on their proposed approach. All in all, we conclude the traditional approach of 50% - 30% is the correct way of evaluating drug-placebo difference. The possible methods to control impact of placebo effect are briefly discussed at the end of this article.

Section 0.11: Optimal Group Sequential Designs Constrained on both Overall and Stage One Error Rates

Optimized group sequential designs proposed in the literature have the aim of minimizing average sample size (ASN) with respect to a prior distribution of treatment effect with overall type I and type II error rates well-controlled. The optimized asymmetric group sequential designs that we present here additionally consider constraints on stopping probabilities at stage one: probability of stopping for futility at stage one when no drug effect exists as well as the probability of rejection when the maximum effect size is true at stage one so that accountability of group sequential design is ensured from the very first stage throughout. As well, non-binding efficacy bounds are used to account for overrunning in common real trials. The shape parameters for Wang-Tsiatis upper bounds and Kim-DeMets lower bounds are utilized to find optimized group sequential designs minimizing ASN while maintaining error and power requirements overall and at stage one. From examples illustrated, the maximum sample size determined through such optimization is smaller than prior optimized designs using other optimization criteria.

Section 0.12: A Two-stage Adaptive Design with a New Combination Test

Inspired by Bauer and Kohne (1994), a method applying Fisher's combination rule to form a two-stage adaptive procedure, together with Box and Muller (1958, referred to as 'BM'), one of the most popular methods of generating standard normal random variable using two independent uniform (0, 1) deviates, a new method (denoted as 'BM combination test') is proposed here to combine two p-values from two disjoint samples for designing a trial with two stages. Procedure is defined with carefully consideration of controlling overall type I error rate under null hypothesis. Operational characteristics including power and expected sample size under both null and alternative hypotheses are investigated. Simulations are used to confirm type I error control. Comparisons of BM combination test with Fisher's combination test are also investigated.

Reference

P. Bauer and K. Kőhne, 1994. Evaluation of Experiment with Adaptive Interim Analysis. *Biometrics*, 50: 1029-1041.
Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.,* 29:610-611
Chen YF, Yang Y, Hung HMJ, Wang SJ. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemp. Clin. Trials* 2011; 32: 592-604.
Fava, M., Evins, A. E., Dorer, D. J., Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics* 72:115-127.
Joseph A. DiMasi , Henry G. Grabowski, Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* 2016, Volume 47, 20–33.
Kim K, Tsiatis AA. Study duration for clinical trials with survival response and early stopping rule. *Biometrics*. 1990, 46(1): 81-92.
Mehta, C.R., Pocock, S.J. (2011). Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples. *Stat Med,* 30(28):3267-84.
O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35:549-56.
Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1977; 38:153-162.
Rihmer, Z., Gonda, X., Döme1, P., Erdős, P., Ormos, M. and Pani, L. (2011) Novel Approaches to Drug-Placebo Difference Calculation: Evidence from Short-Term Antidepressant Drug-Trials. *Human Psychopharmacology: Clinical and Experimental*, 26, 307-312.
Rosenfeld L, Insulin: discovery and controversy. *Clin. Chem*.2002; 48 (12): 2270–88. PMID 12446492.
Rubinstein LV, Gail MH, Santer TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease* 1981; 34:469-479.
Wang SK, Tsiatis AA, Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987, 43:193-199.
Whitehead, J. 'Predicting the duration of sequential survival studies', *Drug Information Journal* 2001; 35: 1387-1400.

CHAPTER 1

Sensitivity Analyses for Informative Censoring in Survival Data: A Trial Example

(published in *Journal of Biopharmaceutical Statistics* ·March 2016 ·DOI: 10.1080/10543406.2016.1167076)

Abstract: In a controlled clinical trial comparing an experimental drug to a control using time to event analysis, the logrank test is normally used to test against the equality between two survival curves when the proportional hazard rate assumption is held, which of course requires non-informative censoring. The authors used an example from a randomized, double-blind, parallel group, low-dose active controlled study comparing the safety and efficacy of two doses (400 mg/day versus 50 mg/day) of study medication used as monotherapy for the treatment of newly diagnosed or recurrent epilepsy. This analysis imputes the event time of subjects considered to have problematic informative censoring to demonstrate the impact of violations in necessary assumptions, and assesses robustness of the p-value as calculated from imputed data as compared with un-imputed data. Assuming a parametric distribution for time to event, had these subjects resulted in an event in the trial after withdrawal, the expected additional time to event is formulated and calculated using methods developed in this paper. Combining the imputed informative censoring subjects with the remainder of the original data, new p-values are obtained using the log-rank test and compared to the original p-value. KM plots are also compared.

Keywords: Survival data; Informative censoring; Robustness; Sensitivity; Expected time to event.

Section 1.1: Introduction

After being randomized into the double-blind phase until the end of study, subjects can have event, or loss to follow-up (due to loss to contact, subject consent or due to adverse event), or remain event free at the time of study termination. The logrank statistic is used to compare the survival distribution of two samples when censoring is non-informative (i.e., the censoring process is independent of the event process). The test was proposed by Nathan Mantel (1966) and was named as ‘logrank test’ by Richard Peto and Julian Peto (1972). Logrank test statistic is constructed by computing the difference between observed and expected number of events in one of the two groups at each unique observed event time and then adding these differences so that a measure for the overall summary across events time points where there is an event is obtained to evaluate two survival distributions in their entirety. The logrank statistic can also be derived as the score test for the Cox proportional hazard model (Cox, David R, 1972) comparing two groups. Based on efficiency of score test, it is therefore asymptotically equivalent to the likelihood ratio test statistic if the proportional hazard model is held, whereas exponential failure time is a special case of the proportional hazard model.

As noted above, logrank test requires non-informative censoring to ensure independence between censoring mechanism and time to event process. In case this assumption is questionable, the validity of this test to measure superiority of one survival curve over the other will be easily challenged. And therefore robustness of p-value from logrank test in this case has to be assessed via sensitivity analyses. For reviewing submitted clinical trial results to support drug label claims, US FDA published a guidance for pharmaceutical industry titled as “E9 Statistical Principles for Clinical Trials”, which indicated their current thinking on this topic as they claimed in the front page. In E9, it is said that “It is important to evaluate the **robustness** of the results and primary conclusions of the trial.” Robustness refers to “the **sensitivity** of the overall conclusions to various limitations of the data, assumptions, and analytic approaches to data analysis”. A real trial is introduced in Section 1.2, with which problematic informative censoring is shown in final data and could possibly invalidate its p-value interpretation. Section 1.3 describes proposed method following up with strategies for sensitivity analyses and subsequent analysis results in Sections 1.4 and 1.5, respectively; and final discussions on method limitations and other methods in Section 1.6 conclude this paper.

Section 1.2: A Trial Example

The objective of this study was to compare the safety and efficacy of 2 doses of topiramate (referred to as ‘TPM’) as monotherapy in pediatric and adult subjects with newly diagnosed (within 3 months) epilepsy characterized by partial-onset or generalized seizures, or with recurrent epilepsy while off of antiepileptic drugs. To ascertain tolerability and to allow for discontinuation of any baseline antiepileptic drugs therapy, eligible subjects received TPM 25 during a 7-day open treatment phase. Between screening (up to 14 days before study entry) and randomization, subjects were to have no more than 1 seizure. Subjects who experienced significant tolerability relating to safety problems during the open-treatment phase were not eligible for randomization. At the end of open treatment, eligible subjects were randomly assigned to either TPM 50 or TPM 400. Antiepileptic drugs therapies, if any, were tapered off prior to randomization. The double-blind phase comprised 2 periods: titration (up to 42 days) and stabilization (of variable duration); subjects who experienced significant tolerability relating to safety problems during the first 21 days of the double-blind phase were withdrawn from the study. Subjects remained in the double-blind phase until i) the first partial onset seizures or generalized seizures, ii) double-blind phase termination (6 months after the last subject was randomized), or iii) withdrawal for protocol-specified reasons (adverse events, subject choice, or lost to follow-up). The efficacy assessment was based on between-group difference in time to first seizure during the double-blind phase. Subjects or their caregivers recorded the date and type of each seizure that occurred in their seizure diaries. A seizure required clinical verification by the investigator. Upon experiencing a seizure, each subject was to contact the investigator, who then evaluated the event in terms of consistency with epileptic partial onset or generalized tonic-clonic seizures.

A total of 487 subjects were enrolled; of those, 16 withdrew during the open treatment phase. Of the 471 subjects randomized, 470 had at least 1 study visit after randomization and were included in the intent-to-treat analysis. Primary efficacy analysis was based on a survival analysis of the difference between TPM 400 and TPM 50 with respect to time to first partial onset seizures or generalized seizures during the double-blind phase (excluding taper). Kaplan-Meier (referred to as 'KM') estimates were calculated for time to first seizure. Statistical significance of the treatment effect was tested by the log-rank test. Trial registration identifier for this study is NCT00231556 at clinicaltrials.gov and trial results were published at *Journal of Child Neurology* (Glauser et al. 2007).

Table 1a lists the completion/withdrawal status along with p-value of efficacy results for original observed data. The first subject's randomization occurred at 19NOV1999; and afterwards eligible patients were continuously randomized until 15AUG2001. There are 470 subjects (TPM 50=234 and TPM400=236), with 90 (38%) and 49 (21%) events occurred in the TPM 50 and TPM 400, respectively. Comparison of the KM survival curves of time to first seizure favored TPM 400 over TPM 50 (p=0.0002; 2-sided log-rank test). When the trial ended at 26FEB2002, there were 217 (TPM 50=105, TPM 400=112) remained event-free at the time of study termination, which were considered as being administratively censored since censoring was caused by trial operation and thus was also considered as non-informative censoring. The proportions of withdrawals due to lost to follow-up and other reason were almost the identical between high and low dose levels (that is: non-differential between two treatment groups), which hinted the claim of non-informative nature for these two kinds of withdrawals. However, at the time of study termination, in the TPM 50 group, 6% (N=13) of subjects had early withdrawal due to adverse event and 4% (N=9) of subjects due to subject choice while having 17% (N=40) of withdrawals due to adverse event and 6% (N=13) of withdrawals due to subject choice in the TPM 400 group. These two types of withdrawals are differential between two treatment groups. Combining these two types of withdrawals together, dis-proportionality in early withdrawal rates between two groups (TPM 400=23% vs. TPM 50=10%) makes people believe that these withdrawals might have informative censoring with being informative with respect to treatment assignment, resulting in violating of non-informative censoring assumption in application of logrank test.

To address this issue, one proposal from US FDA (Food and Drug Administration) reviewer then was to impute informative censoring subjects and treat them as they have had an event occurred at the time of early withdrawal (Table 1.1b). The number of events then becomes 112 (48%) in the TPM 50 group and 102 (43%) in the TPM 400 group, resulting in a big decrease in the difference in event proportion between two groups (5% in difference: TPM 50=48% vs. TPM 400=43%) in this naïve data as compared with original data (17% in difference: TPM 50=38% vs. TPM 400=21%). More importantly, p-value of log-rank test from the naïve data becomes 0.3859 (Table

1.1b), which fails to support the claim of superiority of TPM 400 over TPM 50 in preventing time to first seizure in the double-blind phase. The naïve data are very artificial and incorrect because we only know that subjects who were informatively censored at their withdrawal time but with no knowledge on whether or when event occurred afterwards. Surely for them, there was no event occurring at their date of early withdrawal. From this perspective, the naïve data can be viewed as the 'worst-case- scenario' imputation of the original data. One question to ask next is: what else imputations could possibly depict intermediate scenarios?

Table 1(Tab. 1.1): results from the original data (Table 1.1a) and results from the naive data (Table 1.1b)

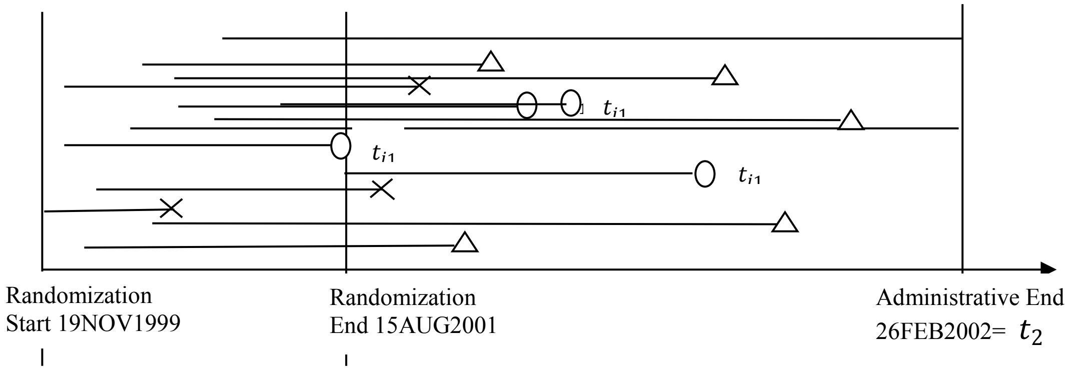
Table 1.1: results from the original data (Table 1.1a) and results from the naive data (Table 1.1b)

Table 1.1a:					
category	Sub-category	TPM 50 N= 234	TPM 400 N= 236	Total N=470	
		n(%)	n(%)	n(%)	p-value =0.0002
Event	seizure	90(38)	49(21)	139(30)	
Informative censoring	Withdrawal due to adverse event	13(6)	40(17)	53(11)	
	Withdrawal due to subject choice	9(4)	13(6)	22(5)	
Non-informative censoring	Administrative censoring	105(45)	112(47)	217(46)	
	Withdrawal due to lost to follow-up	9(4)	10(4)	19(4)	
	Withdrawal due to other reason	8(3)	12(5)	20(4)	
Table 1.1b:					
category	Sub-category	TPM 50 N= 234	TPM 400 N= 236	Total N=470	
		n(%)	n(%)	n(%)	p-value =0.3859
Event	seizure	90(38)	49(21)	139(30)	
	Withdrawal due to adverse event	13(6)	40(17)	53(11)	
	Withdrawal due to subject choice	9(4)	13(6)	22(5)	
Non-informative censoring	Administrative censoring	105(45)	112(47)	217(46)	
	Withdrawal due to lost to follow-up	9(4)	10(4)	19(4)	
	Withdrawal due to other reason	8(3)	12(5)	20(4)	

Section 1.3: Methodology

From analysis of naïve data, we understand that testing of superiority of higher dose versus lower dose via logrank statistic will become non-significant once we consider those informative censoring subjects as event subjects because the test is driven by events and this action adds 53 events to TPM 400 whilst only 22 events to TPM 50, resulting in diluting superiority of TPM 400 over TPM 50 on preventing time to seizure after randomization. To further check sensitivity of p-value in this direction, we propose a method that still assumes that those informative censoring subjects have had an event, but on the contrast, admitting of the event time being later than the withdrawal date, to be consistent with the fact that those subjects didn't have an event at their withdrawal time in the observed data. In Figure 1.1, the upper graph depicts subject's status in the observed data; and after imputation, informative censoring subjects will result in an event between respective withdrawal time and the trial end date 26FEB2002 (see the lower graph in Figure 1.1). The time from randomization to event for informative censoring subjects is imputed with expected additional time to event after being informatively censored at t_{i1} plus observed time course in the double-blind phase (i.e., t_{i1}), had this (ith) subject resulted in first seizure event between withdrawal time t_{i1} and end date t_2. In the upper graph of Figure 1.1, triangle symbol at right end means subjects who had an event in the original data. Subjects with an across symbol at the right end withdrew early due to non-informative reasons (loss to follow-up or other reason). Circled subjects are the ones who are assumed to have informative censoring in the observed data. In the lower graph, suspicious informative censoring subjects are imputed to have an event before or on 26FEB2002, with the long-dash line in bold after their respective early withdrawal time t_{i1} indicating the expected additional time to seizure. Therefore, after imputation, this cohort of subjects will have time to first seizure as total length from randomization time to the predicted event time between early withdrawal time t_{i1} and administrative trial end time 26FEB2002.

Methodology developed below will only apply to informatively censored subjects in the original data. Let X_{ij} and $W_{ij}, j{=}C$ for TPM 50 and E for TPM 400, represent the random variable of time from randomization to first seizure event and from randomization to the time of being censored, respectively, for the ith subject in the jth group who was randomized at time r_{ij} . As explained in the Appendix 1.1, in order to calculate the expected additional time to event for informative censoring subjects, we firstly have to obtain the probability of having an event in $(t_{i1}, t_2]$ given that this subject is event-free at t_{i1}. For a specific event distribution, parameters are estimated from treatment-specific original data with a parametric event distribution imposed (Tables 1.3a-1.3f).



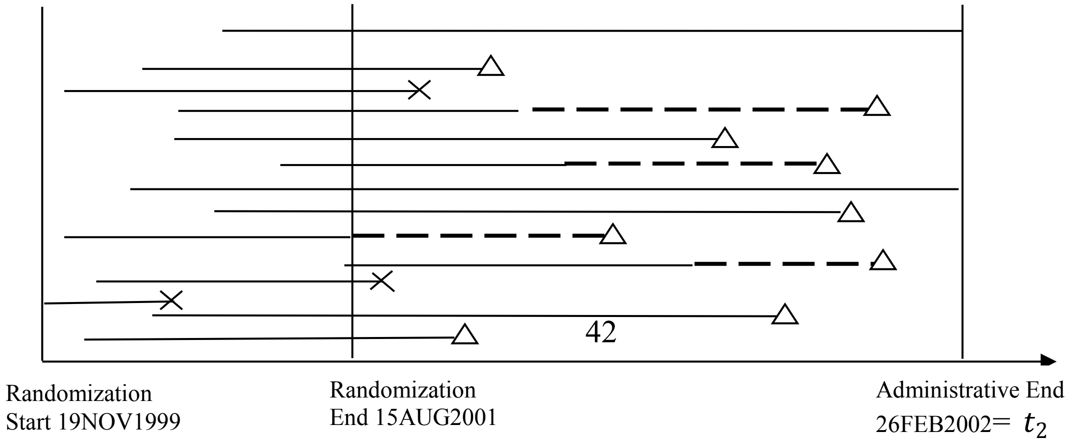


Figure 1(Fig. 1.1): Depiction of imputing process

Figure 1.1: Depiction of imputing process, with the triangle symbol indicating experiencing an event (including events in original data and imputed events in the lower graph), the circle symbol indicating having an informative censoring at t_{i1} in the original data (see in the upper graph), and the across symbol indicating non-informative censoring in the original data, solid line for observed time course and long-dashed line in bold for the expected additional time to event prior to or on the target time t_2. The upper graph represents un-imputed data and the lower graph represents data after imputation.

Based on data from non-informative censoring subjects (i.e., subjects who withdrew due to loss to follow-up or some other reason in this trial), parameters for time to censoring is estimated by: make these non-informative censoring subjects as having an event in the original dataset and the remainder of subjects are all censored. Extract estimated hazard rate parameter by imposing exponential distribution on these created 'event' of time to non-informative censoring. ϕ_E=0.000267784

and ϕ_C =0.0003303452 (Tables 1.3b and 1.3d) are the estimated exponential hazard rates for time to censoring for TPM 400 and TPM 50, respectively.

From Appendix 1.1, it is known that the probability of having an event in $(t_{i1}, t_2]$ for a subject in TPM 50 group in the presence of exponential censoring process competing with event process, given that this subject is event-free at t_{i1} can be expressed as:

$$P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_{i1} - r_{iC}, W_{iC} > t_{i1} - r_{iC})$$

$$= \int_{t_{i1}-r_{iC}}^{t_2-r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} * \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_{i1} - r_{iC})]} dx_{iC}$$

Utilizing independence between event process and exponential censoring process in $(t_{i1}, t_2]$, the above probability can be decomposed to be the product of two components in the integrand, and then the integration is carried out from lower limit

$t_{i1} - r_{iC}$ to upper limit $t_2 - r_{iC}$. The first component $\frac{dP(X_{iC} \le t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})}$ is the derivative of conditional probability of having an event in $(t_{i1}, t_2]$ without competitive censoring (i.e., $P(X_{iC} \le t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})$) with respect to $t_2 - r_{iC}$; and the second component is the conditional exponential censoring survival function $\frac{exp(-\emptyset_C x_{iC})}{exp[-\emptyset_C(t_{i1} - r_{iC})]}$, given this subject is censoring-free at withdrawal time t_{i1}. The expected additional time to event, have this informatively exponential censored subject had resulted in an event in $(\mathrm{t}_{i1}, \mathrm{t}_2]$ is then:

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}} \frac{x_{iC} * \frac{dP(X_{iC} \le t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} * \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_{i1} - r_{iC})]}}{P(X_{iC} \le t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_{i1} - r_{iC}, W_{iC} > t_{i1} - r_{iC})} \, dx_{iC}$$

While other censoring distribution can also plays a role here, as in Equation $1.4'$ from Appendix 1.1, with Weibull censoring, this expected additional time to event is then:

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}} \frac{x_{iC} * \frac{dP(X_{iC} \le t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} * \frac{\omega_C \beta_C x_{iC}{}^{\omega_C - 1} exp(-\beta_C x_{iC}{}^{\omega_C})}{exp(-\beta_C (t_{i1} - r_{iC})^{\omega_C})}}{P(X_{iC} \le t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_{i1} - r_{iC}, W_{iC} > t_{i1} - r_{iC})} \, dx_{iC}$$

with β_C =0.0134838899 and ω_E= 0.6153282175 ($\beta_E = 0.0066766023$ and ω_E =0.6255320211) as parameters estimates for informative Weibull censoring (Tables 1.3e-1.3f)

When censoring process is not essential in calculating expected additional time to event for imputed informative censoring subjects, conditional survival function for censoring process will be dropped from the numerator. And the denominator for probability of having an event in $(t_{i1}, t_2]$, given that this subject is event-free at t_{i1}, can then be expressed as $P(X_{iC} \le t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC},)$ without involving the censoring variable W_{iC}. Therefore, the expected additional time to event for those informative censoring subjects is now:

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}} \frac{x_{iC} * \frac{dP(X_{iC} \leq t_2-r_{iC} | X_{iC} > t_{i1}-r_{iC})}{d(t_2-r_{iC})}}{P(X_{iC} \leq t_2-r_{iC} | X_{iC} > t_{i1}-r_{iC},)} \, dx_{iC}$$

In this section, the algorithm of calculating expected additional time to seizure (bold long-dash line in the lower graph of Figure 1.1) is provided for either with or without considering competitive censoring process. As above, every imputed informative censoring subject will have an event in $(t_{i1}, t_2]$ with the length of time to event equal to sum of time to early withdrawal in the original data (i.e., t_{i1}) and the expected additional time to event in $(t_{i1}, t_2]$, given that this subject was still at risk at t_{i1}. When calculating this expected additional time to event without considering censoring process competing with event process, the integrand part is different from the case with considering it in both denominator and nominator and hence resulting in different expected additional time to event in $(t_{i1}, t_2]$.

Section 1.4: Strategies for Sensitivity Analyses

To make explanations easier, the event distribution and informative censoring distribution (if needed) are both exponential for purpose of illustrating strategies for a series of sensitivity analyses. Figure 1.2 graphically depicts the proposed sensitivity analyses as well as original analysis and naïve analysis proposed by US FDA. In original analysis (referred to as 'O' in Figure 1.2) contains old seizure events data (TPM 50=90 and TPM 400=49 in Table 1.1a), informative censoring subjects whose censoring are probably related to treatment and non-informative censoring subjects whose censoring are considered to be random and independent of treatment assignment. Hazard rates λ_C and λ_E are estimated from original data after imposing a parametric distribution on event time whilst hazard rates of censoring ϕ_C and ϕ_E are estimated using the method mentioned above by 'inverting' original data with non-informative censoring data as 'event' and all the remainders as censoring subjects. Sensitivity strategy S1 in Figure 1.2 denotes the one proposed by US FDA to have all informative censoring subjects have a seizure event at their withdrawal time. Sensitivity analysis strategies S2 and S3 are newly proposed from this paper, in which all or half of the informative censoring subjects will have a seizure event at the predicted time point after withdrawal. Conditional on the fact that informative censoring subjects were still at risk at withdrawal time t_{i1}, the expected time to seizure prior to t_2 is calculated for each informative censoring subject and then the newly created data for this cohort will be added back to the remainder of original data so that p-value and KM plot can be regenerated. As 50%

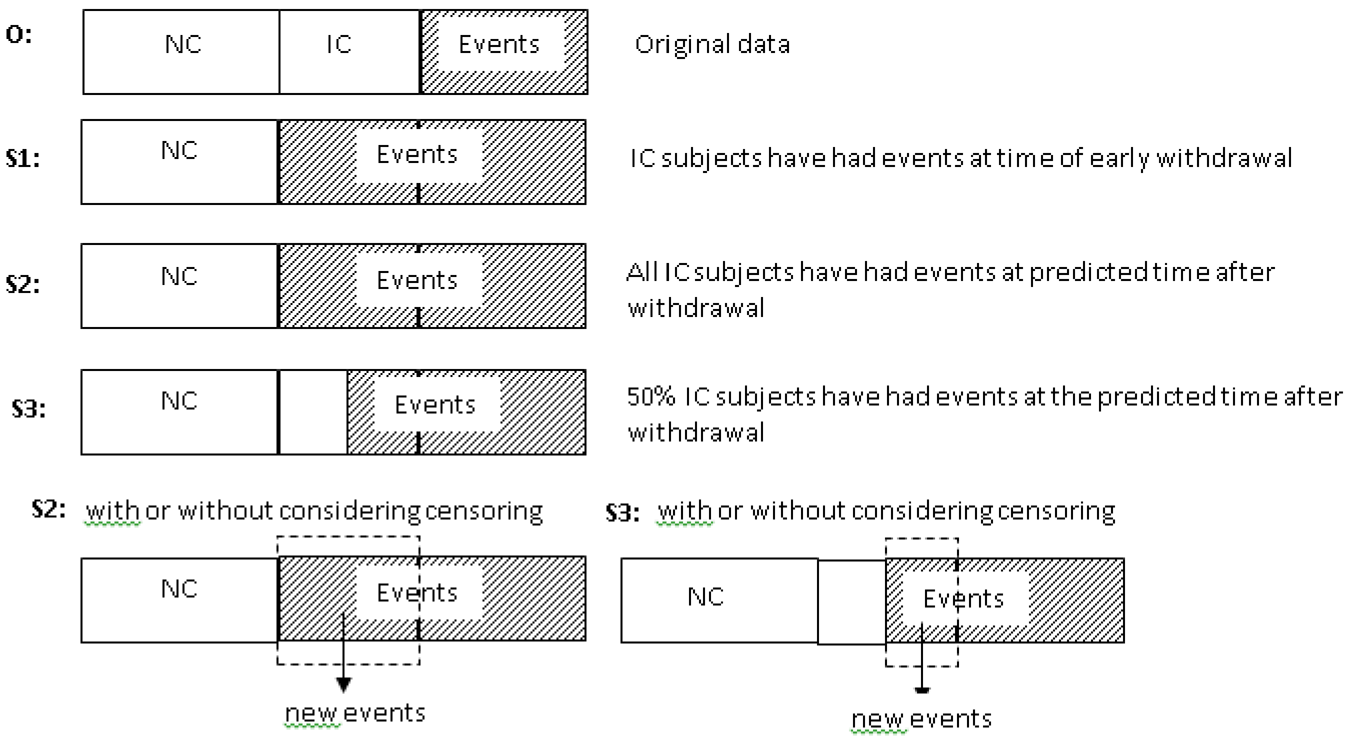

O:
NC
IC
Events
Original data
S1:
NC
Events
IC subjects have had events at time of early withdrawal
S2:
NC
Events
All IC subjects have had events at predicted time after withdrawal
S3:
NC
Events
50% IC subjects have had events at the predicted time after withdrawal
S2: with or without considering censoring
NC
Events
new events
S3: with or without considering censoring
NC
Events
new events

Figure 2(Fig. 1.2): Sensitivity analyses strategies

Figure 1.2: Sensitivity analyses strategies. IC and NC denote informative censoring and non-informative censoring subjects, respectively.

of the informative censoring subjects imputed is because withdrawals due to adverse event or subject choice are generally independent of treatment assignment in normal clinical trials. Therefore, we can’t always assume all informative censoring subjects in this cohort had informative censoring. Of note, regardless of with or without considering censoring, full imputation will have all informative censoring subjects result in an event in $(t_{i1}, t_2]$ and 50% imputing will have half of informative censoring subjects result in an event in $(t_{i1}, t_2]$, while as shown in Section 1.3 and the Appendix 1.1, absence of censoring will change value of integrand when doing integration and thus will result in different expected additional time to event as compared with the case in the presence of censoring.

Section 1.5: Analysis Results

After extracting parameters from original data, for each informative censoring subject, probability of having an event before t_2 is calculated, which is then to be put in the denominator of the integrand in order to obtain the expected additional time to event, had this subject have an event in $(t_{i1}, t_2]$. After imputing those informative censoring subjects, they are put back together with the remainder of original data to do hypothesis testing. Now event/censoring status for intent-to-treat subjects are as represented as in Table 1.2a. From p-value of 0.0002 from original data to 0.3859 with na ïve data, it seems more events added-in, the less significant p-value the test will end up with. To test this speculation, we’ve tried 50% imputation (Table 1.2b). For each informative censoring subject (N=22 in TPM 50, N=53 in TPM 400), one uniform random variable in a range of [0, 1] is generated. This subject will be imputed to have event at his/her expected time before t_2 if this uniform random variable is great than or equal to 0.5, otherwise this subject will still be censored at his/her withdrawal time and no imputation will be conducted. After this manipulation, we created a population with nearly 50% of informative censoring subjects imputed. Results from data with 50% imputation are displayed in Table 1.2b. Of those, 12 out of 22 informative censoring subjects in the TPM 50 group are imputed and 25 out of 53 informative censoring subjects in the TPM 400 are imputed.

Table 2(Tab. 1.2): results from fully imputed data (Table 1.2a) and results from data with 50% imputation (Table 1.2b)

Table 1.2: Results from fully imputed data (Table 1.2a) and results from data with 50% imputation (Table 1.2b)

Table 1. 2a:

category	Sub-category	TPM 50 N= 234	TPM 400 N= 236	Total N=470
		n(%)	n(%)	n(%)
Event	seizure	90(38)	49(21)	139(30)
	Withdrawal due to adverse event (fully imputed)	13(6)	40(17)	53(11)
	Withdrawal due to subject choice (fully imputed)	9(4)	13(6)	22(5)
Non-informative censoring	Administrative censoring	105(45)	112(47)	217(46)
	Withdrawal due to lost to follow-up	9(4)	10(4)	19(4)
	Withdrawal due to other reason	8(3)	12(5)	20(4)

Table 1.2b:

category	Sub-category	TPM 50 N= 234	TPM 400 N= 236	Total N=470
		n(%)	n(%)	n(%)
Event	seizure	90(38)	49(21)	139(30)
	Withdrawal due to adverse event (imputed)	6(3)	15(6)	21(4)
	Withdrawal due to subject choice (imputed)	6(3)	10(4)	16(3)
Non-informative censoring	Withdrawal due to adverse event	7(3)	25(11)	32(7)
	Withdrawal due to subject choice	3(1)	3(1)	6(1)
	Administrative censoring	105(45)	112(47)	217(46)
	Withdrawal due to lost to follow-up	9(4)	10(4)	19(4)
	Withdrawal due to other reason	8(3)	12(5)	20(4)

To understand how informative censoring subjects could potentially impact final summary measure of p-value from logrank test due to violation of independent censoring assumption in the original data, we investigate imputations under the scenarios: different parametric event distribution, with/without considering censoring, fully imputed or only with 50% imputation, and with or without treatment-specific parameters reverted:

i) p-values from logrank tests with data imputation for informative censoring subjects without considering of censoring process competing with the event process (Table 1.3a),

ii) the same as i) but with considering exponential censoring in calculating expected addition time to event (Table 1.3b),

iii) The same as i) but with treatment-specific parameters swapped (Table 1.3c),

iv) Without considering censoring and with treatment-specific parameters swapped (Table 1.3d),

where, as noted in Section 1.4, parameters swap/reverted refers to switch the set of estimated parameters for time to event by arm, plus switch those for time to informative dropout by arm.

Tables 1.3a-1.3d have shown p-values of imputations with intermediate states in-between the original and the naïve data. New methods are developed to address informative censoring issue while making use of the fact that those subjects were not yet having had an event at their withdrawal time. When all these 77 subjects are imputed (Row 3 in Tables 1.3a-1.3d), p-values become at 0.1 level regardless of event distribution type, ranging from 0.1165 to 0.1687. The extent of p-values is consistent among different parametric event distributions. Calculation of the expected additional time to seizure makes use of group information by extracting treatment-specific parameters as well as subject-level information by having subject specific conditional density conditional upon the fact of being at risk at withdrawal time. p-values at 0.1 level for full imputation show that original p-value of 0.0002 is quite robust because events added to TPM400 group from imputation is more than two times higher than that of TPM50 group (e.g., 53 vs. 21) so that imputation in this case indeed introduced a great extent of dilution to the overall effect on preventing from time to seizure between high and low dose groups.

To see the variants for this worst case scenario (‘worse’ means resulting in a decrease in treatment effect after imputation), imputation to calculate expected additional time to event is also conducted while considering censoring process accompanying with the event process (Table 1.3b), it is good to see that the p-values are still at 0.1 level. The impact of competing censoring process has little impact on conditional probability of having an event prior to the trial end date and hence has little impact on the expected length of having an event in $(t_{i1}, t_2]$, given that this subject is event-free at t_{i1}.

To check the worse situation of each of the above imputed strategies, we inverted two sets of parameters when calculating the expected additional time to event for chosen informative censoring, making the estimated parameters from TPM 50 group (or TPM 400) to do imputation for TPM 400 (or TPM 50) IC subjects so that we can further dilute treatment difference between TPM 400 and TPM 50, because, for this cohort of imputed informative censoring subjects, treatment effect is in the opposite direction of the overall effect in the whole intent-to-treat analysis set. Results are shown in Tables 1.3c and 1.3d, which are uniformly worse than (as expected) their counterparts in Tables 1.3a-1.3b, but in a small extent. This, per our opinion, further supports our conclusion that impacts from this set of informative censoring subjects on original p-value are not substantial. All cases with treatment-specific parameters reverted has a larger p-value than that of its counterpart without purposely inverting in Tables 1.3a and 1.3b, but the excess level is 0.02 or less for fully imputed cases and only 0.002 level or less for 50% imputed cases.

Tables 3e and 3f are added per reviewer's suggestion to assess impact of different distribution of censoring on robustness of p-values after imputation. Comparing with exponential censoring, Weibull censoring results in a little bigger p-value for every parametric event distribution without/with parameter swap (Table 1.3e vs. Table 1.3b and Table 1.3f vs. Table 1.3d), whereas general conclusions above regarding robustness of p-value after imputation with/out censoring and with/out parameters swap remain the same.

Figure 1.3 graphically depicts all p-values in Tables 1.3a-1.3f into one graph to illustrate the whole picture of our imputation strategy, with left-most as p-value from the original data, right-most as p-value from the naïve data, 50% as well as full imputation as intermediate imputations proposed in this paper. Significance decreases from left for being most significant, still significant for all 50% imputations irrespective to with or without competitive censoring process and parameter swap between to comparing groups, non-significant for full imputations, and the most non-significant case for p-value is computed from the naïve data.

Figures 1.4 contains Kaplan-Meier plots for some proposed cases of imputation against both KM plot from original data as well as the naïve data, because KM plot is an alternative way to show the differences among different imputations. The biggest separation between two groups occurs in original data (the upper left in Figure 1.4) and no separation is shown in naïve data (the upper right in Figure 1.4). Separations between two groups are bigger in the plots with 50% imputation than those with full imputation, regardless of distribution assumption and whether the parameter set being swapped or not. Note that the same set of KM plots for other parametric distributions are done but not shown in this paper due to space limitation.

Table 3(Tab. 1.3): p-values from logrank tests

Table 1.3: p-values from logrank tests with imputations for informative censoring subjects (fully imputed or with 50% imputation) when calculation of the expected additional time to event is in the absence of censoring (1.3a), exponential censoring is present in (1.3b), in the absence of censoring together with parameter swap (1.3c), in the presence of exponential censoring together with parameter swap (1.3d), in the presence of Weibull censoring (1.3e) and in the presence of Weibull censoring together with parameter swap (1.3f)

Table 1.3a: in the absence of censoring				
	Exponential	Weibull	Log normal	Log logistic
Parameters	λ_C=0.0013703428 λ_E=0.0007085123	α_c = 0.0100808764 γ_c = 0.6609148602 α_E =0.0047444152 γ_E= 0.6791830664	μ_c=6.4815452017 σ_c=2.2883766324 μ_E=7.7589738974 σ_E=2.5726080137	α_c = 0.007403468 γ_c = 0.7639210804 α_E = 0.0039600968 γ_E= 0.7365174685
p-value(full)	0.1165	0.1367	0.1441	0.1362
p-value(50% imputation)	0.0106	0.0117	0.0119	0.0115
Table 1.3b: in the presence of exponential censoring				
	Exponential	Weibull	Log normal	Log logistic
Parameters	As in Table 1.3a	As in Table 1.3a	As in Table 1.3a	As in Table 1.3a
	ϕ_C=0.000267784, ϕ_E =0.0003303452			
p-value(full imputation)	0.1207	0.1383	0.1496	0.1420
p-value(50% imputation)	0.0109	0.0116	0.0121	0.0118
Table 1.3c: in the absence of censoring and with parameter swap				
	Exponential	Weibull	Log normal	Log logistic
Parameters	λ_E=0.0013703428 λ_C=0.0007085123	α_E = 0.0100808764 γ_E = 0.6609148602 α_C =0.0047444152 γ_C= 0.6791830664	μ_E=6.4815452017 σ_E=2.2883766324 μ_C=7.7589738974 σ_C=2.5726080137	α_E = 0.007403468 γ_E = 0.7639210804 α_C = 0.0039600968 γ_C= 0.7365174685
p-value(full imputation)	0.1394	0.1565	0.1663	0.1658
p-value(50% imputation)	0.0123	0.0126	0.0129	0.0129
Table 1.3d: in the presence of exponential censoring and with parameter swap				
	Exponential	Weibull	Log normal	Log logistic
Parameters	As in Table 1.3c	As in Table 1.3c	As in Table 1.3c	As in Table 1.3c
	ϕ_E=0.000267784, ϕ_C =0.0003303452			
p-value(full imputation)	0.1429	0.1584	0.1687	0.1652
p-value(50% imputation)	0.0120	0.0125	0.0130	0.0129
Table 1.3e: in the presence of Weibull censoring				
	Exponential	Weibull	Log normal	Log logistic
Parameters	As in Table 1.3a	As in Table 1.3a	As in Table 1.3a	As in Table 1.3a
	β_E = 0.0066766023, ω_E =0.6255320211 and β_C =0.0134838899, ω_E= 0.6153282175			
p-value(full imputation)	0.144	0.1589	0.1638	0.1575
p-value(50% imputation)	0.0119	0.0125	0.0125	0.0123
Table 1.3f: in the presence of Weibull censoring and with parameter swap				
	Exponential	Weibull	Log normal	Log logistic
Parameters	As in Table 1.3c	As in Table 1.3c	As in Table 1.3c	As in Table 1.3c
	β_E =0.0134838899, ω_E= 0.6153282175 and β_C = 0.0066766023, ω_C =0.6255320211			
p-value(full imputation)	0.187	0.2005	0.2082	0.2074
p-value(50% imputation)	0.0139	0.144	0.0147	0.0147

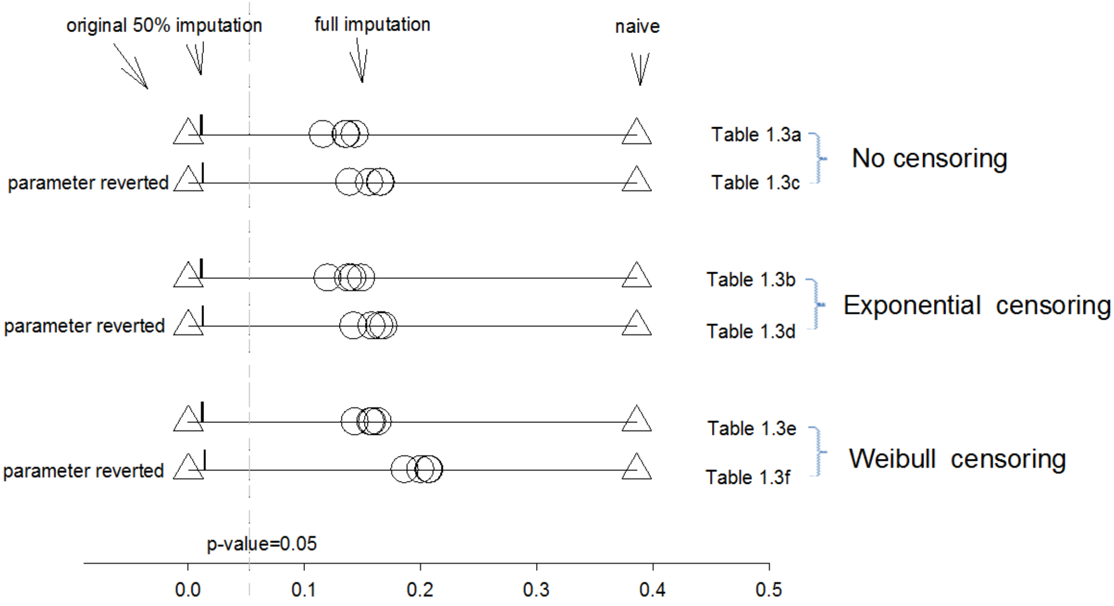


Figure 3(Fig. 1.3): P-value summary for sensitivity analyses

Figure 1.3: P-value summary for sensitivity analyses in Tables 1.3a-1.3f. From left to right, left triangle indicates p-value from original data, following up four vertical bars at 0.01 level and four circles between 0.1 and 0.21 represent p-values obtained from exponential, Weibull, log normal and log logistic event distribution, and the right triangle indicates p-value from the na ïve data.

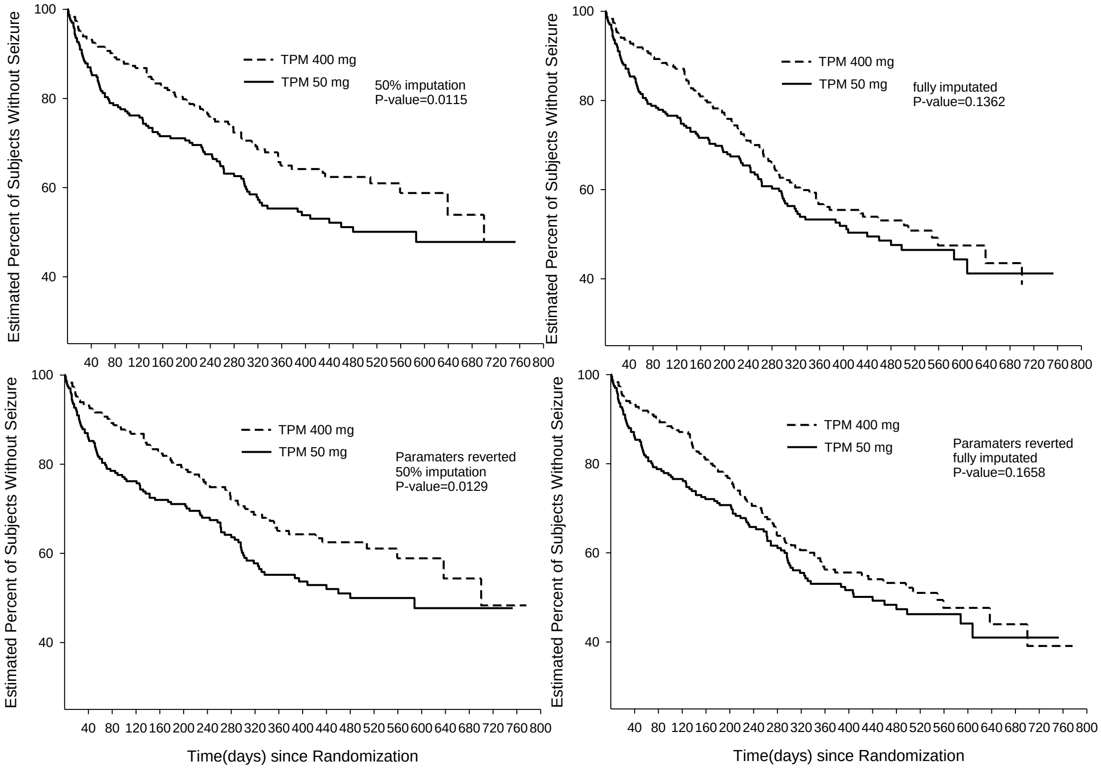


Figure 4(Fig. 1.4): KM plots

Figure 1.4: KM plots for: original data (upper left), na ïve data (upper right), informative censoring subjects exponentially distributed without considering exponential censoring in calculating expected additional time to event (with only 50% imputation: lower left; fully imputed: lower right).

Section 1.6: Discussion

Starting from a real example for a clinical trial with survival endpoints accompanying with obvious informative censoring, authors develop methods to do sensitivity analyses to demonstrate the robustness of p-value from logrank test. It is to estimate treatment-specific parameters for each group after imposing a particular parametric distribution; then calculate subject specific probability of having an event in $(t_{i1}, t_2]$, given that this subject is event-free at t_{i1} with or without considering censoring process competing with event process. Proposed imputations using expected time to event plus original time course as the event time for imputed informative-censoring subjects resulted in p-values at 0.01+ or 0.1+ level for exponential censoring and a little higher for Weibull censoring, regardless of parametric event distribution, with or without considering censoring, even additionally with treatment-specific parameters swapped between groups.

To think of these imputations from a different angle (also see in Figure 1.3), the original data resulted in a strong claim in significance regarding treatment effect for comparing high dose with the lower dose on time to seizure. Results from partial imputations (50% imputation conducted here) are deemed to be the most reasonable ones among all methods mentioned in this paper. The reasoning should be as the following. As noted in the primary paper for this study (Glauser et al., 2007), "*The most common adverse events, excluding typical childhood illnesses, were headache, appetite decrease, weight loss, somnolence, dizziness, concentration/attention difficulty, and paresthesia*.". Fifty-three subjects who withdrew early due to adverse events, although with differential dropout rates between groups, shouldn't be all considered as informative censoring and relating to study medication due to the nature of these events. This supports the usage of 50% imputation rather than full imputation. Therefore, p-values with 50% imputation are around 0.01 for both exponential and Weibull censoring, as compared with p-value 0.3859 from the naive data, further corroborating the significance claim from the original data.

Parameter swap can further dilute treatment effect as treatment effect within this small group of imputed informative-censoring subjects is intentionally reversed and is in the opposite direction of the overall effect. However, p-values only increase by 0.002 or less (Table 1.3a vs. Table 1.3c and Table 1.3b vs. Table 1.3d) as compared with 50% imputation without parameter swap, irrespective of parametric distributions and irrespective of being in the presence or in the absence of censoring. Along this road, all doubtful withdrawals due to adverse events or subject choice are imputed (i.e., fully imputed) assuming all subjects in these two categories being informatively censored and they are all assumed to have had an event in $(t_{i1}, t_2]$, which is of course an extremely strong assumption as in this case none of the adverse events and subject-choice withdrawals is assumed not to be related to treatment assignment. p-values now become 0.11 - 0.2082, non-significant but still much less than 0.3859, the one from the na ïve data. For now, we take back what we said early in Section 1.1 about that the imputation done in the naïve data is the 'worst-case scenario' imputation for this trial data. To our opinion, p-values with full imputation, instead of the p-value from the na ïve data, should serve as the worst-case scenario among all proper imputations for this trial data, because in the na ïve data all informative-censoring subjects are assumed to have had an event occurring right at their withdrawal time point and this is something definitely not true. Therefore, p-values with proposed full imputation, 0.11-0.2082, rather than 0.3859 (i.e., the one from the na ïve data) should serve as the upper bound for p-values from sensitivity analyses after taking account of the variability introduced by violating independent censoring assumption.

The whole set of exercises have done two things here: 1) provide a method for sensitivity analysis, and 2) confirm the robustness of p-value of log-rank test for the original data. In order to think of how these sensitivity analyses corroborate p-value from original data, we can imagine other hypothetical results with a different p-value profile: for example, if p-values from 50% imputation already reach out to a non-significance level of 0.05, then the robustness of original p-value under this case will be fiercely challenged comparing with what have been observed in Tables 1.3a-1.3f and Figure 1.3. Anyway, statistical methods proposed in this paper together with proposed analysis strategies could possibly help trial statisticians conduct sensitivity analyses in facing trials with a similar issue.

There is a rich literature on publications of sensitivity analyses for informative censoring in survival trials. Among them, the method of inverse probability-of-censoring weights (referred to as ‘IPCW’) (Robins and Finkelstein 2000) has been considered as the most popular one for now, whilst at the same time being criticized by its limitations (Howe et.al. 2011). Our method is a supplement to available ones, which is much easier to digest by clinical statisticians as not being associated with behind scene martingale theories and it is very easy to implement. Due to limited time, IPCW method hasn’t been investigated by the author yet but comparison of methods will be the next thing to investigate.

Trial Registration clinicaltrials.gov Identifier: NCT00231556

References:

Cox DR (1972), "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B* 34 (2): 187–220.

Howe JC, Cole JS, Chmiel JS and Munoz A (2011), “Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias”. *Am J. Epidemiology*, 173(5): 569-577.

Johnson & Johnson Pharmaceutical Research & Development Clinical study Report, EDMS-USRA-7548071:3.0. (2002) “A Randomized, Double-Blind, Parallel Group, Monotherapy Study to Compare the Safety and Efficacy of Two Doses of Topiramate in the Treatment of Newly Diagnosed or Recurrent Epilepsy”.

Mantel N (1966), "Evaluation of survival data and two new rank order statistics arising in its consideration". *Cancer Chemotherapy Reports*, 50 (3): 163–70.

Robins JM, Finkelstein DM (2000), “Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests”. *Biometrics*, 56(3):779–788.

Peto R and Peto J (1972), "Asymptotically Efficient Rank Invariant Test Procedures". *Journal of the Royal Statistical Society, Series A* (Blackwell Publishing) 135 (2): 185–207.

Glauser TA, Dlugos DJ, Dodson WE, Grinspan A, Wang S, Wu SC (2007), “Topiramate Monotherapy in Newly Diagnosed Epilepsy in Children and Adolscents”. *Journal of Child Neurology* 22 (6); 693-699.

Statistical Appendix 1.1:

Notations used are re-stated here to ensure completeness of this appendix and methodology is described using the control group as an example. To impute informative censoring subjects, let X_{ij} and $W_{ij},\ j = C, E$, represent random variables of time to event and time to censoring for ith subject treated with control (C or TPM 50) and treatment (E or TPM 400) medications, respectively. All calculations in treatment group will be defined similarly. For the ith subject in the control group TPM 50, r_{iC} and t_{i1} are the randomization date and the date of informative censoring (e.g., withdrawal due to adverse event or subject choice in this trial), respectively. Let t_{i2} be the time of administrative trial end date 26Feb2002, which is date that the last patient had end-of-study visit performed. As t_{i2} is the same for all subjects across two groups, we denote $\mathrm{t_{i2}}$ as t_2 in this paper. Subscript i however can't be omitted in $r_{iC},\ r_{iE}$ and t_{i1}, as they are subject-level randomization dates and subject-level informative censoring date. It is known that the event time for subject i will be at least $t_{i1} - r_{iC}$ due to early withdrawal at time t_{i1}. Assumed that this subject had resulted in an event between t_{i1} and t_2, the first quantity to be calculated is the probability of having an event in $(t_{i1}, t_2]$, given that this subject is event-free at t_{i1}. Next, we return to our objective of calculating: Had this subject resulted in an event prior to t_2, what would it be for the expected additional time of having this event after t_{i1} and prior to t_2? Before calculating the expected additional time to event for each imputed informative censoring subject, let's calculate probability of having an event in$(t_{i1}, t_2]$, given that subject is event-free at t_{i1}, which is needed for calculation of expected additional time to event in Step 2) below.

Step 1): For these informative censoring subjects, probability of having an event in $(t_{i1}, t_2]$ when there is an independent censoring process competes with event process is:

$$P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_{i1} - r_{iC}, W_{iC} > t_{i1} - r_{iC})$$

$$= E_{X_{iC}} \left[I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC}) P(x_{iC} < W_{iC} | W_{iC} > t_{i1} - r_{iC}) \right] \quad (1.1)$$

$$= E_{X_{iC}} \left[\frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_{i1} - r_{iC})]} \right] \quad (1.2)$$

$$= \int_{t_{i1} - r_{iC}}^{t_2 - r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} * \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_{i1} - r_{iC})]} dx_{iC} \quad (1.3)$$

Equation 1 is based on independence of time to censoring (i.e., W_{iC}) and event process (i.e., X_{iC}). Equation 1.2 makes use of time to non-informative censoring, which is exponentially distributed with hazard rate $\emptyset_C$. $\frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_{i1} - r_{iC})]}$ is the conditional exponential survival function for time to censoring, given that subject still in the risk set at time t_{i1}. $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})$ is the probability of having an event in $(t_{i1}, t_2]$ in the absence of censoring, given that the subject is still in the risk set at time t_{i1}. In order to calculate conditional probability of having an event in the presence of censoring, one component in the integral is taking derivative of conditional probability in the absence of censoring with respect to $t_2 - r_{iC}$. That is $\frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})}$ and the second component is the conditional exponential survival function of the censoring variable (See in Equation 1.3).

Step 2): The expected time to event, had this informative censoring subject resulted in an event in $(t_{i1}, t_2]$ is:

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}} \frac{x_{iC} * \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})}{d(t_2 - r_{iC})} * \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C (t_{i1} - r_{iC})]}}{P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_{i1} - r_{iC}, W_{iC} > t_{i1} - r_{iC})} \, dx_{iC} \qquad (1.4)$$

where the probability calculated in Equation 1.3 is now the denominator of the integrand in Equation 1.4 . To understand the above formulation, one way is to think of P(A|B)=P(AB)/P(B). P(B) is the conditional probability of have an event in $(\mathrm{t}_{\mathrm{i}1}, \mathrm{t}_2]$ for informative censoring subjects in the presence of censoring. For different parametric time to event distributions, density of event time (i.e., $f_{X_{iC}}(t)$, row 1 in Table 1.4), is used to obtain conditional probability of having an event in $(t_{i1}, t_2]$, which is $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_{i1} - r_{iC})$ (row 2 of Table 1.4). Subsequently, after taking derivative with respect to the random variable $t_2 - r_{iC}$ (row 3 in Table 1.4), conditional probability of having an event in $(t_{i1}, t_2]$, in the presence of censoring as in Equation 1.3 or row 4 of Table 1.4 will be calculated for different parametric event distributions. Finally, the expected time to event in $(t_{i1}, t_2]$ can be calculated, had this informative censoring subject resulted in an event before or on t_2.

In case of non-exponential censoring, other conditional survival density of time to censoring, which is the component of ($P(x_{iC} < W_{iC} | W_{iC} > t_{i1} - r_{iC})$ in Equation 1.1, will be plugged in Equations 1.2, 1.3 and 1.4 in order to calculate the expected time to event in $(t_{i1}, t_2]$ for imputed subject i. For example, in case time to censoring having Weibull distribution with parameters of β_c and ω_c, time to censoring density function then becomes $\omega_c \beta_c x_{iC}{}^{\omega_c - 1} exp(-\beta_c x_{iC}{}^{\omega_c})$ and survival function at time $t_{i1} - r_{iC}$ is $exp(-\beta_c (t_{i1} - r_{iC})^{\omega_c})$, resulting in conditional survival density being $P(x_{iC} < W_{iC} | W_{iC} > t_{i1} - r_{iC}) = \frac{\omega_c \beta_c x_{iC}{}^{\omega_c - 1} exp(-\beta_c x_{iC}{}^{\omega_c})}{exp(-\beta_c (t_{i1} - r_{iC})^{\omega_c})}$.

Therefore Equations 1.2, 1.3, 1.4 will become Equations $1.2', 1.3', 1.4'$ respectively as follows.

$$E_{X_{iC}}\left[\frac{dP(X_{iC}\le t_2-r_{iC}|X_{iC}>t_{i1}-r_{iC})}{d(t_2-r_{iC})}*\frac{\omega_c\beta_c x_{iC}{}^{\omega_c-1}\,exp(-\beta_c x_{iC}{}^{\omega_c})}{exp(-\beta_c(t_{i1}-r_{iC})^{\omega_c})}\right] \qquad (1.2')$$

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}}\frac{dP(X_{iC}\le t_2-r_{iC}|X_{iC}>t_{i1}-r_{iC})}{d(t_2-r_{iC})}*\frac{\omega_c\beta_c x_{iC}{}^{\omega_c-1}\,exp(-\beta_c x_{iC}{}^{\omega_c})}{exp(-\beta_c(t_{i1}-r_{iC})^{\omega_c})}dx_{iC} \qquad (1.3')$$

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}}\frac{x_{iC}*\frac{dP(X_{iC}\le t_2-r_{iC}|X_{iC}>t_{i1}-r_{iC})}{d(t_2-r_{iC})}*\frac{\omega_c\beta_c x_{iC}{}^{\omega_c-1}\,exp(-\beta_c x_{iC}{}^{\omega_c})}{exp(-\beta_c(t_{i1}-r_{iC})^{\omega_c})}}{P(X_{iC}\le t_2-r_{iC},X_{iC}<W_{iC}|X_{iC}>t_{i1}-r_{iC},\ W_{iC}>t_{i1}-r_{iC})}\ dx_{iC} \qquad (1.4')$$

And the rest for calculating expected additional time for imputed subjects remains the same as case of exponential time to censoring illustrated in Steps 1 and 2.

Calculation will be much simplified if there is no censoring process in competition with event process. Without considering censoring, the expected length time of being an event in $(t_{i1}, t_2]$ for this informative censoring subject is then degenerated to:

$$\int_{t_{i1}-r_{iC}}^{t_2-r_{iC}}\frac{x_{iC}*\frac{dP(X_{iC}\le t_2-r_{iC}|X_{iC}>t_{i1}-r_{iC})}{d(t_2-r_{iC})}}{P(X_{iC}\le t_2-r_{iC}|X_{iC}>t_{i1}-r_{iC},)}\ dx_{iC} \qquad (1.5)$$

The numerator of integrand is x_{iC} times quantity from row 3 in Table 1.4 for respective parametric event distribution and the denominator is the conditional probability calculated in row 2 of Table 1.4. Table 1.4 contains necessary ingredients for computation, in which rows 3 is used in the numerator of integrand for both cases with or without considering censoring and row 2 and row 4 are used in the denominator part of the integrand for the case in the absence of censoring and the case in the presence of censoring, respectively.

Table 4(Tab. 1.4): Ingredients for calculation of the expected additional time to event

Table 1.4: Ingredients for calculation of the expected additional time to event after withdrawal when parametric event distributions are exponential, Weibull, log normal and log logistic, respectively. Row 1, 2 and 3 display density of event distribution, conditional probability of having an event in $(t_{i1}, t_2]$ in the absence of censoring and conditional density of having an event in $(t_{i1}, t_2]$ in the absence of censoring, respectively. Row 3 is the first integrand component in calculating Row 4, which is the conditional probability of having an event in $(t_{i1}, t_2]$ in the presence of exponential censoring. Row 5 is in the counterpart of Row 4 but with Weibull censoring.

Event Distribution		exponential	Weibull
Row 1	$f_{X_{ic}}(t)$	$\lambda_c \exp(-\lambda_c t)$	$\gamma_c \alpha_c t^{\gamma_c - 1} \exp(-\alpha_c t^{\gamma_c})$ where $\sigma_c = 1/\gamma_c$ and $\alpha_c = \exp(-\mu_c/\sigma_c)$
Row 2	$P(X_{ic} \le t_2 - r_{ic} \| X_{ic} > t_{i1} - r_{ic})$	$1 - \frac{\exp[-\lambda_c(t_2 - r_{ic})]}{\exp[-\lambda_c(t_{i1} - r_{ic})]}$	$1 - \frac{\exp[-\alpha_c(t_2 - r_{ic})^{\gamma_c}]}{\exp[-\alpha_c(t_{i1} - r_{ic})^{\gamma_c}]}$
Row 3	$\frac{dP(X_{ic} \le t_2 - r_{ic} \| X_{ic} > t_{i1} - r_{ic})}{d(t_2 - r_{ic})}$	$\frac{\lambda_c exp[-\lambda_c(t_2 - r_{ic})]}{exp[-\lambda_c(t_{i1} - r_{ic})]}$	$\frac{\alpha_c(t_2 - r_{ic})^{\gamma_c - 1} exp[-\alpha_c(t_2 - r_{ic})^{\gamma_c}]}{exp[-\alpha_c(t_{i1} - r_{ic})^{\gamma_c}]}$
Row 4	$P(X_{ic} \le t_2 - r_{ic}, X_{ic} < W_{ic} \| X_{ic} > t_{i1} - r_{ic}, W_{ic} > t_{i1} - r_{ic})$ in the presence of exponential time to censoring	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{\lambda_c\, exp[-\lambda_c(t_2 - r_{ic})]}{\exp[-\lambda_c(t_{i1} - r_{ic})]} * \frac{exp(-\emptyset_c x_{ic})}{exp[-\emptyset_c(t_{i1} - r_{ic})]} dx_{ic}$	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{\alpha_c(t_2 - r_{ic})^{\gamma_c - 1} exp[-\alpha_c(t_2 - r_{ic})^{\gamma_c}]}{exp[-\alpha_c(t_{i1} - r_{ic})^{\gamma_c}]} * \frac{exp(-\emptyset_c x_{ic})}{exp[-\emptyset_c(t_{i1} - r_{ic})]} dx_{ic}$
Row 5	$P(X_{ic} \le t_2 - r_{ic}, X_{ic} < W_{ic} \| X_{ic} > t_{i1} - r_{ic}, W_{ic} > t_{i1} - r_{ic})$ in the presence of Weibull time to censoring	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{\lambda_c\, exp[-\lambda_c(t_2 - r_{ic})]}{exp[-\lambda_c(t_1 - r_{ic})]} * \frac{\omega_c \beta_c {x_{ic}}^{\omega_c - 1} exp(-\beta_c {x_{ic}}^{\omega_c})}{exp(-\beta_c(t_{i1} - r_{ic})^{\omega_c})} dx_{ic}$	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{\alpha_c(t_2 - r_{ic})^{\gamma_c - 1} exp[-\alpha_c(t_2 - r_{ic})^{\gamma_c}]}{exp[-\alpha_c(t_{i1} - r_{ic})^{\gamma_c}]} * \frac{\omega_c \beta_c {x_{ic}}^{\omega_c - 1} exp(-\beta_c {x_{ic}}^{\omega_c})}{exp(-\beta_c(t_{i1} - r_{ic})^{\omega_c})} dx_{ic}$

Event Distribution		log normal	Log logistic
Row 1	$f_{X_{ic}}(t)$	$\frac{1}{\sqrt{2\pi\sigma_c t}} \exp(-\frac{1}{2}(\frac{\log(t) - \mu_c}{\sigma_c}))$	$\frac{\alpha_c \gamma_c t^{\gamma_c - 1}}{(1 + \alpha_c t \gamma_c)^2}$ where $\gamma_c = 1/\sigma_c$ and $\alpha_c = \exp(-\mu_c/\sigma_c)$
Row 2	$P(X_{ic} \le t_2 - r_{ic} \| X_{ic} > t_{i1} - r_{ic})$	$1 - \frac{1 - \Phi(\frac{\log(t_2 - r_{ic}) - \mu_c}{\sigma_c})}{1 - \Phi(\frac{\log(t_{i1} - r_{ic}) - \mu_c}{\sigma_c})}$	$1 - \frac{1 + \alpha_c(t_{i1} - r_{ic})^{\gamma_c}}{1 + \alpha_c(t_2 - r_{ic})^{\gamma_c}}$
Row 3	$\frac{dP(X_{ic} \le t_2 - r_{ic} \| X_{ic} > t_{i1} - r_{ic})}{d(t_2 - r_{ic})}$	$\frac{exp(-\left(\frac{log(t_2 - r_{ic}) - \mu_c}{\sigma_c}\right)^2)/(2 * \sqrt{2\pi})}{1 - \Phi\left(\frac{log(t_{i1} - r_{ic}) - \mu_c}{\sigma_c}\right)} (\frac{1}{\sigma_c} * \frac{1}{t_2 - r_{ic}})$	$\frac{[1 + \alpha_c(t_{i1} - r_{ic})^{\gamma_c}] * \gamma_c \alpha_c (t_2 - r_{ic})^{\gamma_c - 1}}{[1 + \alpha_c(t_2 - r_{ic})^{\gamma_c}]^2}$
Row 4	$P(X_{ic} \le t_2 - r_{ic}, X_{ic} < W_{ic} \| X_{ic} > t_{i1} - r_{ic}, W_{ic} > t_{i1} - r_{ic})$ in the presence of exponential time to censoring	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{exp \frac{-\left(\frac{log(x_{ic}) - \mu_c}{\sigma_c}\right)^2}{2 * \sqrt{2\pi}}}{1 - \Phi\left(\frac{log(t_{i1} - r_{ic}) - \mu_c}{\sigma_c}\right)} \left(\frac{1}{\sigma_c} * \frac{1}{t_2 - r_{ic}}\right) * \frac{exp(-\emptyset_c x_{ic})}{exp[-\emptyset_c(t_{i1} - r_{ic})]} dx_{ic}$	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{[1 + \alpha_c(t_{i1} - r_{ic})^{\gamma_c}] * \gamma_c \alpha_c {x_{ic}}^{\gamma_c - 1}}{[1 + \alpha_c {x_{ic}}^{\gamma_c}]^2} * \frac{exp(-\emptyset_c x_{ic})}{exp[-\emptyset_c(t_{i1} - r_{ic})]} dx_{ic}$
Row 5	$P(X_{ic} \le t_2 - r_{ic}, X_{ic} < W_{ic} \| X_{ic} > t_{i1} - r_{ic}, W_{ic} > t_{i1} - r_{ic})$ in the presence of Weibull time to censoring	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{exp \frac{-\left(\frac{log(x_{ic}) - \mu_c}{\sigma_c}\right)^2}{2 * \sqrt{2\pi}}}{1 - \Phi\left(\frac{log(t_{i1} - r_{ic}) - \mu_c}{\sigma_c}\right)} \left(\frac{1}{\sigma_c} * \frac{1}{t_2 - r_{ic}}\right) * \frac{\omega_c \beta_c {x_{ic}}^{\omega_c - 1} exp(-\beta_c {x_{ic}}^{\omega_c})}{exp(-\beta_c(t_{i1} - r_{ic})^{\omega_c})} dx_{ic}$	$\int_{t_{i1} - r_{ic}}^{t_2 - r_{ic}} \frac{[1 + \alpha_c(t_{i1} - r_{ic})^{\gamma_c}] * \gamma_c \alpha_c {x_{ic}}^{\gamma_c - 1}}{[1 + \alpha_c {x_{ic}}^{\gamma_c}]^2} * \frac{\omega_c \beta_c {x_{ic}}^{\omega_c - 1} exp(-\beta_c {x_{ic}}^{\omega_c})}{exp(-\beta_c(t_{i1} - r_{ic})^{\omega_c})} dx_{ic}$

CHAPTER 2

Sample Size Increase during a Survival Trial When Interim Results are Promising

(published at *Communication in Statistics: theory and method* **DOI:** 10.1080/03610926.2015.1137596)

Abstract: This paper is to extend Mehta and Pocock (2010) to provide a way in doing sample size increase in survival trials. Sample space is divided by observed test statistic at interim into three zones: unfavorable, promising and favorable, within which sample size (required number of events) has a proper increase if falling into the promising zone and otherwise remains unchanged. Simulations with scenarios in the presence/absence of censoring, with/without adaptation, and allowing 4 folds vs. 2-folds of increase in sample size are compared.

Keyword: Survival Trials; Promising Zone; Sample Size Re-estimation; Group Sequential Design.

Section 2.1: Introduction

Clinical trials to fulfil the requirements of new drug application need to show both efficacy in a disease indication and safety for patients who have been exposed to investigational drug for a long enough time period. Comparing time to event for experimental drug against the control group, log-rank test is normally used to test against the equality between two survival curves when proportional hazard assumption is held. An anticipated log hazard ratio (control vs. experimental) is assumed prior to trial start in order to design a trial ensuring desired power to detect treatment difference when a certain amount of relative superiority indeed exists. However, design adaptations (i.e., with respect to either increase in sample size, drop treatment arms/doses, change entry criteria, change randomization ratio, even change endpoint or other areas) are imperative especially when the trial is in an underexplored territory regarding unmet medical needs. In a seminar talk held in 2010 (http://catalyst.harvard.edu/docs/biostatsseminar/Pocock_04_March_2010.pdf), some trial examples were mentioned on how trial adaptations could possibly rescue a failure trial in drug development history in several disease areas. Here is one related to survival analysis. The

Cardiac Insufficiency Bisoprolol Study (CIBIS) began at 1989 to answer the question “Does bisoprolol reduce mortality in heart failure”. With an underpowered design, 641 subjects with chronic heart failure of various etiologies and a left ventricular ejection fraction <40% entered into the double-blind phase (bisoprolol=320 and placebo=321). Mean duration in the double-blind phase was 1.9 years. Equivalent withdrawal rates in the double-blind phase occurred between two groups (82 on placebo and 75 on bisoprolol). P-value of 0.22 from log-rank test failed to show the superiority of bisoprolol over placebo in reducing the mortality in heart failure (hazard ratio: 0.80; 95% confidence interval: 0.56 to 1.15); and 67 patients died on placebo and 53 on bisoprolol (CIBS, 1994). CIBS-II trial was conducted to re-check the effect of decreasing all-cause mortality in chronic heart failure. Results were published in The Lancet (CIBS-II, 1999), with which 2647 symptomatic patients from Europe were enrolled and randomly assigned to 1.25 mg bisporolol (N=1327) and placebo (N=1320) daily. CIBIS-II was stopped early after the second interim analysis because bisoprolol showed a significant benefit in all-cause mortality over placebo (P-value<0.0001; hazard ratio=0.66; and 95% confidence interval 0.54-0.81). There was significantly less all-cause mortality among patients on bisoprolol than those on placebo (156 [11.8%] vs 228[17.3%]). The estimated annual mortality rate from CIBS-II was 8.8% in the bisoprolol group and 13.2% in the placebo group. It took almost ten years from failing an under-powered study CIBS-I to a successful re-testing of the same hypothesis in CIBS-II. And eventually drug approval was obtained in 1999. Have sample size adaptation had been implemented in CIBS-I trial, would CIBS-II trial be no longer needed? This will be answered in Section 2.4 as an illustration example for the proposed method. Because modifying ongoing phase III trial designs seems a contradictory action against its confirmatory nature and any adaptation during the trial could potentially jeopardize trial’s integrity and inflate false positive rate of the trial, the PhRMA Adaptive Working Group published a White Paper concerning operational issues (PhRMA, 2007), while the FDA more conservatively adopted an attitude to wait for more experience on sample size re-estimation based on unblinded treatment information (FDA, 2010).

Among many methodological articles on sample size re-estimation, a focus has been on how to preserve the overall type I error rate. A circular conditional error was proposed and an adjusted critical value for final analysis based on power requirement while preserving type I error rate for normal data was proposed by Proschan and Hunsberger (1995). Cui, Hung and Wang (1999) proposed combining the Wald statistic from two stages using pre-specified weights, in which weighted Wald statistic under null hypothesis is normally distributed with mean zero and variance of one resulting from independence from statistic before and after interim analysis. Bauer and Kohne (1994) proposed using Fisher's combination test to combine two p-values from stage one and two in order to control type I error rate. Another way proposed by Lehmacher and Wassmer (1999) is to use inverse normal function. Above methods to combine independent test statistic or p-values from independent cohorts of subjects are easily applied for normal data and binary data since subjects to be included prior to or after the interim are naturally in different cohorts and inherently independent in terms of endpoint measuring clinical benefits. Survival data are different in which subjects who are ongoing at the time of interim analysis (i.e., administratively censored) will definitely contribute to the final analysis in a way either being censored or experiencing an event upon final analysis. In controlling type I error rate in adaptive designs, Muller and Schafer (2001) generalized methods for controlling overall type I error rate and showed that the overall type I error rate can be preserved unconditionally for any possible adaptation, provided that the conditional error based interim test statistic would have been obtained had there been no adaptation is preserved.

All adaptive methods discussed above are to use non-standard final test statistic in which subjects enrolled before or after the interim analysis are treated (or weighted) differently. This stimulated a hot discussion on the appropriateness of assigning different weights to subjects enrolled before or after interim adaptation. A seemly more attractive way is to stick on conventional statistic without a weighting strategy while using accumulative data upon study termination and unadjusted critical value for final decision with which it then seems violation of “one patient one vote” principle introduced by unequal weights is avoided. Chen, Demets and Lan (2004) took an initial step in this direction and showed that type I error rate won’t get inflated using conventional final analysis and unadjusted critical value if the interim results are located in a “promising zone”. Next, Gao, Ware and Mehta (2008) worked out the statistical rational for Chen, Demets and Lan (2004) and further expanded the range of the promising zone based on conditional power using treatment effect observed at interim analysis. Mehta and Pocock (2010) extended Chen, Demets and Lan (2004) a bit in a more practical manner by tabulating explicit cutoff value for the promising zone determined by pre-specified information vector, ratio of maximum sample size relative to pre-planned sample size, and observed test statistic at interim.

This paper starts with historical clinical trials of CIBS-I and CIBS-II in Section 2.1 to address the importance of having sample size increase for clinical trials with survival data. Section 2.2 describes trial hypothesis in testing equality of two survival curves using conventional log-rank test and weighted log-rank test after sample size increase at interim. Section 2.3 extends the method proposed by Mehta and Pocock (2010) to survival data, emphasizing on obtaining sample space based on interim test statistic divided as unfavorable, promising and favorable zones before trial starts. Section 2.4 revisits two CIBS trials and calculates the required sample size for stage two after interim analysis, had proposed sample size re-estimation algorithm been implemented in CIBS-I trial. Section 2.5 includes extensive simulations on exponential survival data: 1) in the presence or absence of censoring; 2) sample size increase occurred in the middle of the trial, in the early part or in the later part of the trial; and 3) ratio of total maximum sample size after adaptation relative to the planned total sample size being large (i.e., dmax/d=4) or moderate (i.e., dmax/d=2). Section 2.6 summarizes all the findings and discusses possible refinements in future research.

Section 2.2: Log-rank and Weighted Log-rank

Assuming time to failure for control subjects is exponentially distributed with a constant hazard of λ_c, the median time of $M_c = \ln(2)/\lambda_c$, to test against null hypothesis of equal survival curves, i.e., $\ln(\Delta) = 0$, where $\Delta = \frac{\lambda_c}{\lambda_E}$, λ_E being the hazard rate for experimental group subjects, one wishes to have a pre-specified power in testing one-sided alternative of $\ln(\Delta) > 0$ (or $\Delta > 1$) against $\ln(\Delta) = 0$. During the double-blind phase, time to failure is independently and identically distributed (i.e., i.i.d.) within a treatment group and independent of subject's entry time as well as independent of time to censoring, where time to censoring are i.i.d.s with expo(ϕ), with the same hazard rate of time to censoring for subjects in two comparative groups. Let $\widehat{\Delta}$ be the estimator of Δ. The reason to use $\ln(\Delta)$ instead of Δ is because $\ln(\widehat{\Delta})$ is less skewed and has a more accurate asymptotic approximation. With exponential distribution, hazard function is constant, which is actually not necessary for logrank statistic. Logrank statistic can also be derived as the score test for the Cox Proportional Hazard model (Cox, David R, 1972) comparing two groups only requiring proportional hazard (i.e., constant hazard ratio instead of constant hazard rate). Based on efficiency of the score test, it is therefore asymptotically equivalent to the likelihood ratio test statistic if the proportional hazard model holds, whereas exponential failure time is a special case of it. For a fixed sample design, to test $H_0\colon \ln(\Delta) = 0$ vs. $H_A\colon \ln(\Delta) > 0$ at one-sided significance level of $\alpha/2$ and power of $1 - \beta$ under alternative hypothesis, one needs to link log hazard ratio with type I and II error requirements using asymptotical properties of logrank statistic; and then calculate the required number of evens to ensure testing power when alternative hypothesis is true. For a group sequential design, a coefficient is to be multiplied with the requirement number of events calculated for corresponding fixed sample design to account for multiple testing over stages (Jennison and Turnbull, 2000).

Without loss of generality, one considers a two-stage group sequential design with upper efficacy boundary vector $\{b_1, b\}$ and the number of events vector $\{d_1, d\}$ with subscript 1 indicating analysis at interim. The corresponding information vector is $\{t_1, 1\}$ with $t_1 = \frac{d_1}{d}$. Without adaptation, interim will occur when d_1 events are accumulated and final analysis will occur when d events are accumulated with corresponding log-rank test statistic Z_1 for interim and Z for final using accumulative data up to analysis time. Null hypothesis of equal hazard rates (or hazard ratio being 1 under proportional hazard) between groups will be rejected if z_1 being greater or equal to critical value b_1; or if not, after adaptation, study continues to accumulate d_2^* number of events. Note that if there is no adaptation when null is not rejected at interim analysis, trial continues to accumulate additional d_2 (i.e.,$d - d_1$) events before final analysis.

Again, when there is a need for sample size adaptation, as in the CIBS I trial, d_2 might be increased to d_2^*. Then simply comparing conventional test statistic Z^*(conducting logrank test using accumulative data) based on d^* ($d^* = d_1 + d_2^*$) with the unadjusted final critical value b to do hypothesis testing at final analysis might inflate type I error. At the time of interim analysis, accumulative data are put together for log-rank test, including subjects who have had an event or experienced censoring prior to interim cutoff date and subjects who are still ongoing will be administratively censored at time of cutoff date. As those administratively censored subjects at interim analysis could either have an event or to be censored at time of final analysis, there is no way to simply use subjects enrolled after interim to do analysis for the independent increment as what we generally do for both normal and binary data. Inspired by Equations 2.3-2.6 in Proschan, Lan, Wittes (2006), we propose using imaginary independent increment X_2^* to obtain weighted log-rank test Z^*. As defined in Proschan, Lan, Wittes (2006), let $B(t_1) = \sqrt{t_1}Z_1$ and $B(1) = Z$ for our two-stage group sequential design. Z is log-rank test statistic with no adaptation in sample size, a function of d_2.

$B(1) = B(t_1) + B(1) - B(t_1)$

$Z = \sqrt{t_1}Z_1 + \sqrt{1 - t_1}\,X_2^*$ because independent increment $B(1) - B(t_1) = \sqrt{1 - t_1}\,X_2^*$

After sample size increase, t_1 becomes $t_1^* = \frac{d_1}{d^*}$. Similarly, we will have

$Z^* = \sqrt{t_1^*}Z_1 + \sqrt{1 - t_1^*}\,X_2^*$, where Z^* is log-rank test statistic after sample size adaptation, a function of d_2^*.

After adaptation, we now get imaginary independent increment $X_2^* = \frac{Z^* - \sqrt{t_1^*}Z_1}{\sqrt{1-t_1^*}}$. Putting X_2^* back into equation for Z, we then have

$Z_{CHW}^* = \sqrt{t_1}Z_1 + \sqrt{1 - t_1}\,X_2^* = \sqrt{t_1}Z_1 + \sqrt{1 - t_1}\frac{Z^* - \sqrt{t_1^*}Z_1}{\sqrt{1-t_1^*}}$.

Because Z_{CHW}^* is a test type similar to the one for normal/binary data in Cui, Hung and Wang (1999), we use subscript ‘CHW’ to indicate it. As noted above, Z_{CHW}^* under null hypothesis shares the same distributional assumptions with Z in absence of adaptation and thus decision rule of $Z_{CHW}^* \geq b$ can be used for final analysis without jeopardizing controlling of type I error rate. However, $Z_{CHW}^{*} \geq b$ is not used in this paper and we indeed try to find a way of using $Z^{*} \geq b$ even after sample size adaptation.

In summary, in $\mathrm{Z} = \sqrt{t_1} Z_1 + \sqrt{1-t_1}\, X_2^{*}$, weight of t_1 is pre-specified, independent of observed Z_1 and independent of imaginary increment X_2^{*}. Plugging X_2^{*} (obtained from $Z_{CHW}^{*} = \sqrt{t_1^{*}} Z_1 + \sqrt{1-t_1^{*}}\, X_2^{*}$) into Z helps creating a weighted log-rank test statistic Z_{CHW}^{*}, which is a function of t_1^{*} and hence a function of d_2^{*} as well, but having the same distributional property as Z to control type I error rate.

Another component in need is the conditional power assuming current trend being carried towards the end of the trial. That is:

$P_{HA}(Z^{*} \geq b | Z_1 = z_1, \hat{\theta} = \ln(\hat{\Delta}))$, where $\ln\big(\hat{\Delta}\big) = \frac{z_1}{\sqrt{\frac{d_1}{4}}}$ and assumes the trend observed at interim is carried forward to the final analysis. Equation Z^{*} then becomes $\sqrt{t_1^{*}} z_1 + \sqrt{1-t_1^{*}}\, X_2^{*}$ after observing $Z_1 = z_1$. We now have conditional power as:

$$P_{HA}\left(\frac{Z^{*}\sqrt{\frac{d_1+d_2^{*}}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2^{*}}{4}} * \sqrt{\frac{d_2^{*}}{4}} * \hat{\theta}}{\sqrt{\frac{d_2^{*}}{4}}} \geq \frac{b\sqrt{\frac{d_1+d_2^{*}}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2^{*}}{4}} * \sqrt{\frac{d_2^{*}}{4}} * \hat{\theta}}{\sqrt{\frac{d_2^{*}}{4}}}\right)$$

$$= P_{HA}\left(\frac{Z^{*}\sqrt{\frac{d_1+d_2^{*}}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2^{*}}{4}} * \sqrt{\frac{d_2^{*}}{4}} * \frac{z_1}{\sqrt{\frac{d_1}{4}}}}{\sqrt{\frac{d_2^{*}}{4}}} \geq \frac{b\sqrt{\frac{d_1+d_2^{*}}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2^{*}}{4}} * \sqrt{\frac{d_2^{*}}{4}} * \frac{z_1}{\sqrt{\frac{d_1}{4}}}}{\sqrt{\frac{d_2^{*}}{4}}}\right)$$

$$= 1 - \Phi\left(\frac{b\sqrt{\frac{d_1+d_2^{*}}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2^{*}}{4}} * \sqrt{\frac{d_2^{*}}{4}} * \frac{z_1}{\sqrt{\frac{d_1}{4}}}}{\sqrt{\frac{d_2^{*}}{4}}}\right) \qquad (2.1)$$

because left-hand side of equation becomes standard normal variable with mean 0 and variance of 1 asymptotically. Obviously, conditional power with current trend is a function of d_2^{*}

additional number of events. After iterative search, we can find d_2^* to ensure conditional power of 1-β that is to have the conditional power the same as the overall power of the trial. When there is no sample size change, $d_2^* = d_2$, then the conditional power carrying current trend becomes:

$$1\text{-}\Phi\left(\frac{b\sqrt{\frac{d_1+d_2}{4}} - z_1\sqrt{\frac{d_1}{4}} - \sqrt{\frac{d_2}{4}} * \sqrt{\frac{d_2}{4}} * \frac{z_1}{\sqrt{\frac{d_1}{4}}}}{\sqrt{\frac{d_2}{4}}}\right) \tag{2.2}$$

Alternatively, there is a closed form for d_2^* to ensure power of 1-β asymptotically, which is actually used for the calculations and simulations in this article. Detailed derivations on this closed form can be obtained upon request from the correspondence author. And the closed form of d_2^* to ensure asymptotic conditional power is as follows:

$$d_2^* = \frac{d_1}{z_1^2}\left\{\frac{b\sqrt{d} - z_1\sqrt{d_1}}{\sqrt{d_2}} + z_{1-\beta}\right\}^2 \tag{2.3}$$

Equation (2.3) is actually the same as Equation 3.11 in Wassmer, G. (2006), provided that there is no stratification plus having 1:1 randomization ratio between treatment and placebo.

Section 2.3: Sample Space of the First-stage Statistic: Unfavorable, Promising and Favorable Zones

Without sample size adaptation, decision using $Z \geq b$ will ensure type I error rate control. With sample adaptation, $Z_{CHW}^* \geq b$ can ensure type I error rate as explained in Section 2.2. In the meanwhile, there are two more ways to control type I error rate: $Z^* \geq b^*$ and $Z^* \geq b$, where the latter is to use both conventional test statistic (but based on d^* after adaptation) and unadjusted critical value b to avoid violating 'one person one vote' concern as mentioned before and it is also what this paper is dedicated to, while with specific interests in applications in survival data. Actually, $Z^* \geq b^*$, as described below, can also control type I error rate. And strategies used in method for using $Z^* \geq b^*$ also plays an important role in developing strategies for the method for using $Z^* \geq b$.

Without futility bound at interim (i.e. no stopping for futility), the unconditional type I error spent at stage two without and with adaptation, respectively, is as follows:

$\int_{-\infty}^{b_1} P_{H0}(Z \geq b|\, Z_1 = z_1)\, \emptyset(z_1)\, dz_1$ and $\int_{-\infty}^{b_1} P_{H0}(Z^* \geq b^*|\, Z_1 = z_1)\, \emptyset(z_1)\, dz_1$

Obviously, in order to control overall type I error rate, we have to have

$P_{H0}(Z \geq b|\, Z_1 = z_1) = P_{H0}(Z^* \geq b^*|\, Z_1 = z_1)$ for all $z_1 \in (-\infty, b_1)$

That is $P_{H0}(Z \geq b|\, Z_1) = P_{H0}(Z^* \geq b^*|\, Z_1)$ unconditionally.

Similar to computing conditional power, getting conditional error is under null effect of hazard ratio being one rather than carrying observed effect towards the end of the trial, therefore, the left-hand side becomes $LHS = 1 - \Phi(\frac{b\sqrt{\frac{d_1+d_2}{4}} - z_1\sqrt{\frac{d_1}{4}}}{\sqrt{\frac{d_2}{4}}})$

and the right-hand side is $RHS = 1 - \Phi(\frac{b^*\sqrt{\frac{d_1+d_2^*}{4}} - z_1\sqrt{\frac{d_1}{4}}}{\sqrt{\frac{d_2^*}{4}}})$

Equating both, we have b^* as the function of b. That is:

$$b^* = \frac{1}{\sqrt{d_1+d_2^*}}\left[\left(\sqrt{\frac{d_2^*}{d_2}}\left(b\sqrt{d_1+d_2} - z_1\sqrt{d_1}\right)\right) + z_1\sqrt{d_1}\right] \tag{2.4}$$

So after adaptation type I error rate will be well-controlled when $Z^* \geq b^*$ is used as the final rejection rule. Above derivation is an implementation of Gao, Ware and Mehta (2008) to survival data. Now, let's go back to the question asked in Section 2.1, is it possible to stick to decision rule using both conventional test after adaptation (i.e., Z^*) and original critical value b while still not inflating type I error rate even with a sample size increase after interim? So the goal here is: instead of controlling type I error rate using $Z \geq b$ without adaptation or $Z^* \geq b^*$ after adaptation, when is it applicable to use $Z^* \geq b$, the conventional test Z^* after adaptation but still with original critical value b? In doing that, weighting strategy which violates of "one patient one vote" is not used at final analyses, which therefore makes communications between statisticians and clinical people much easier. Since b^* (also defined as $b^*(z_1, d_2^*)$) is a function of d_2^* whereby d_2^* is linked to conditional power, one can only do adaptation in a sample space of z_1 where conditional power based on z_1 and d_2^* leads to $b^*(z_1, d_2^*) \le b$. See below for picking up d_2^* in Steps i)-ii).

For these cases of z_1 in a region resulting in $b^*(z_1, d_2^*) \le b$, it can be proved that $P_{H0}(Z^* \ge b^* | Z_1) \ge P_{H0}(Z^* \ge b | Z_1)$. Specifically, because b_2^* is chosen so that $\int_{b_1}^{+\infty} P_{H0}(Z^* \ge b^* | Z_1 = z_1)\emptyset(z_1)dz_1 = \alpha_2$, with α_2 being alpha level spent at stage two after interim, the usage of $Z^* \ge b$ at the final analysis only when z_1 is in the region resulting in $b^*(z_1, d_2^*) \le b$ will always result in a type I error rate at stage 2 being less than or equal to the pre-allocated alpha for stage 2. Mathematically, $= \alpha_1 + \alpha_2 =$

$\Phi(b_1) + \int_{b_1}^{+\infty} P_{H0}(Z^* \ge b^* | Z_1 = z_1)\emptyset(z_1)dz_1 \ge \Phi(b_1) + \int_{b_1}^{+\infty} P_{H0}(Z^* \ge b | Z_1 = z_1)\emptyset(z_1)dz_1$, because during the sample size adaptation d_2^* given z_1 is chosen in a region with $b^*(z_1, d_2^*) \le b$.

Following Mehta and Pocock (2010), here are the steps to do sample size increase during a survival trial when interim results are promising:

i) For each $Z_1 = z_1$, find corresponding $d_2^{\#}$ so that conditional power carrying current trend till the study end being $1 - \beta$.

ii) $d_2^* = \min(d_2^{\#},\ d_{2,max} = d_{max} - d_1)$ to account for budget limit.

iii) For a pair of (z_1, d_2^*), calculate adjusted critical value $b^*(z_1, d_2^*)$ using Equation (2.4).

iv) For a pair of (z_1, d_2^*), calculate new conditional power $CP_{\hat{\theta}}(z_1, d_2^*)$ based on adjusted additional d_2^* events after interim using Equation (2.1).

v) For this particular z_1, calculate original conditional power $CP_{\widehat{\theta}}(z_1, d_2)$ based on planned additional d_2 events after interim using Equation (2.3).

vi) Iterative Steps i)-v) for $z_1 \in [0.01,4.00]$ by increment of 0.01.

Using all values obtained from above in Steps i)-vi), a promising zone is created as follows:

vii) Plotting $b^*(z_1, d_2^*)$ versus z_1.

viii) Plotting the curve of preplanned critical value line b for final analysis which is a horizontal line.

ix) Plotting the curve of conditional power $CP_{\widehat{\theta}}(z_1, d_2)$ against pair of z_1 and pre-planned d_2.

x) Promising zone is defined as: $\mathbb{p} = \{CP_{\widehat{\theta}}(z_1, d_2): b^*(z_1, d_2^*) \leq b\ \}$ and the minimal conditional power is: $CP_{\widehat{\theta},min}$=inf$\{CP_{\widehat{\theta}}(z_1, d_2): b^*(z_1, d_2^*) \leq b\}$.

xi) $CP_{\widehat{\theta},max}$=$\{CP_{\widehat{\theta}}(z_1, d_2): CP_{\widehat{\theta}}(z_1, d_2) = 1 - \beta\}$.

xii) The sample space of z_1 is then divided into three regions:

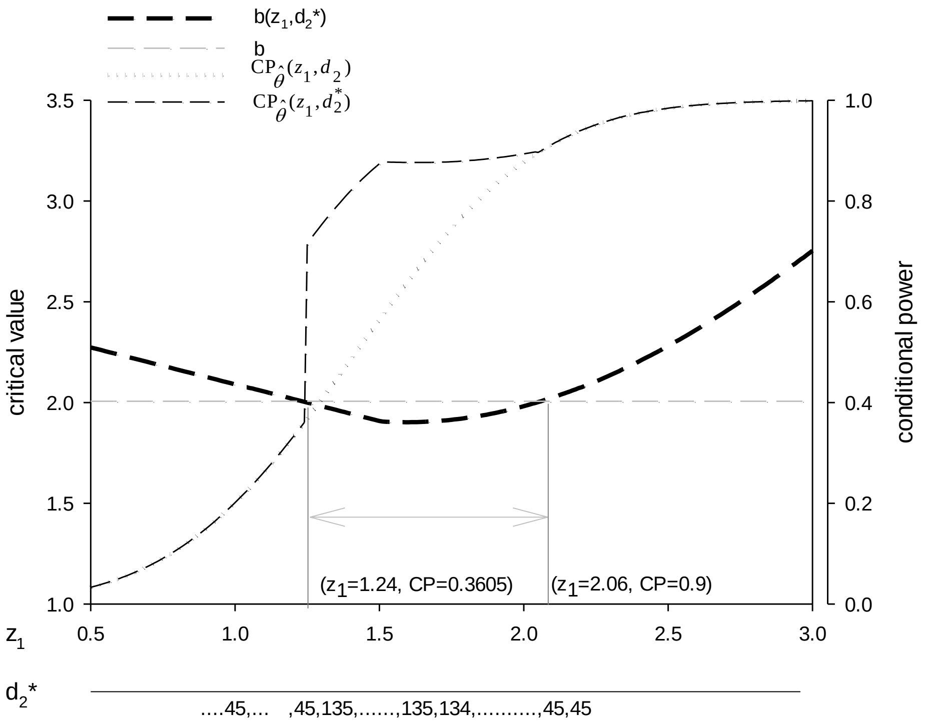
The unfavorable zone $CP_{\widehat{\theta}}(z_1, d_2) \in [0, CP_{\widehat{\theta},min})$

The promising zone $CP_{\widehat{\theta}}(z_1, d_2) \in [CP_{\widehat{\theta},min}, 1 - \beta]$

The favorable zone $CP_{\widehat{\theta}}(z_1, d_2) \in (1 - \beta, 1]$

xiii) Set $\mathrm{d}_2^* = \mathrm{d}_2$ when z_1 is located in both unfavorable and favorable zones.

xiv) Plotting the curve of conditional power $\mathrm{CP}_{\widehat{\theta}}(\mathrm{z}_1, \mathrm{d}_2^*)$ against z_1 based on the adapted d_2^* to check conditional power change after boosting sample size from d_2 to d_2^* in this promising zone.



z_1	d^*_2	z_1	d^*_2	z_1	d^*_2	z_1	d^*_2
1.24	45	**1.45**	**135**	**1.66**	**99**	**1.88**	**64**
1.25	135	**1.46**	**135**	**1.67**	**97**	**1.89**	**63**
1.26	135	**1.47**	**135**	**1.68**	**95**	**1.90**	**61**
1.27	135	**1.48**	**135**	**1.69**	**93**	**1.91**	**60**
1.28	135	**1.49**	**135**	**1.70**	**91**	**1.92**	**59**
1.29	135	**1.50**	**135**	**1.71**	**89**	**1.93**	**58**
1.30	135	**1.51**	**134**	**1.72**	**88**	**1.94**	**57**
1.31	135	**1.52**	**132**	**1.73**	**86**	**1.95**	**56**
1.32	135	**1.53**	**129**	**1.74**	**84**	**1.96**	**55**
1.33	135	**1.54**	**126**	**1.75**	**82**	**1.97**	**54**
1.34	135	**1.55**	**123**	**1.76**	**81**	**1.98**	**52**
1.35	135	**1.56**	**121**	**1.77**	**79**	**1.99**	**51**
1.36	135	**1.57**	**119**	**1.78**	**78**	**2.00**	**50**
1.37	135	**1.58**	**116**	**1.79**	**76**	**2.01**	**50**
1.38	135	**1.59**	**114**	**1.80**	**75**	**2.02**	**49**
1.39	135	**1.60**	**112**	**1.81**	**73**	**2.03**	**48**
1.40	135	**1.61**	**109**	**1.82**	**72**	**2.04**	**47**
1.41	135	**1.62**	**107**	**1.83**	**70**	**2.05**	**46**
1.42	135	**1.63**	**104**	**1.84**	**69**	**2.06**	**45**
1.43	135	**1.64**	**102**	**1.85**	**68**	**2.07**	**45**
1.44	135	**1.65**	**100**	**1.86**	**66**	2.08	45
				1.87	**65**		

Figure 5(Fig. 2.1): Promising zone, adjusted critical value and conditional power

Figure 2.1: Promising zone, adjusted critical value and conditional power curves for a two stage design with WT boundaries with shape parameter of 0.15, t_1=0.5, $\alpha = 0.025, \beta = 0.1$, d_{max}/d=2 and no early stopping for futility.

2.1a: adjusted final critical value $b^*(z_1, d_2^*)$, conditional power based on d_2 and d_2^*

respectively versus z_1.

2.1b: Adjusted d_2^* versus z_1 in the promising zone.

Figure 2.1a graphically represents how optimal zone is chosen based on Steps i)-xiv). A two-stage group sequential design using Wang-Tsiatis (1987) (WT) upper boundaries with shape parameter of 0.15, t_1=0.5, $\alpha = 0.025, \beta = 0.1$, $\mathrm{d}_{\max}$/d=2, hazard ratio $\Delta = 2$ and no early stopping for futility. Then one obtains b vector = (2.556876, 2.006084), required events $\mathrm{d}_1 = 45$, $\mathrm{d}_2 = 45$, d=90, $\mathrm{d}_{\max} = 180$ and $\mathrm{d}_{2,\max} = 135$. For each z_1 in the sample space, $\mathrm{d}_2^{\#}$ is sought out to ensure conditional power $\mathrm{CP}_{\hat{\theta}}(\mathrm{z}_1, \mathrm{d}_2^{\#})$ being 0.9 using Equation (2.1) with assuming observed effect size at interim being carried towards the end of the trial; then the adapted sample size for stage two is $\mathrm{d}_2^* = \min(\mathrm{d}_2^{\#}, \mathrm{d}_{2,\max} = \mathrm{d}_{\max} - \mathrm{d}_1 = 135)$ with truncation from above due to budget limit. Figure 2.1a has the second x-axis below the main x-axis z_1 to show the corresponding d_2^* associated with each sample point of z_1. Adjusted $b^*(z_1, d_2^*)$ per Equation (2.3) is the final adjusted critical value to control type I error rate when using decision rule $Z^* \geq b^*(z_1, d_2^*)$, where Z^* is the conventional log-rank test statistic based on accumulative data upon study termination without weighting strategy. Next, using $Z^* \geq b = 2.006084$ as the rejection rule whenever z_1 is residing in the zone with $b^*(z_1, d_2^*) \leq b$ will control the type I error rate at 0.025 level because probability of conventional test statistic being greater than or equal to $b^*(z_1, d_2^*)$ under null hypothesis is exactly 0.025 and hence resulting in type I error less than or equal to 0.025 when test statistic is compared with b in the promising zone with $b^*(z_1, d_2^*) \leq b$. Black Long-dash line decreases first and then increases in z_1 with an interval being less than equal to the horizontal line of original critical value b, the grey long-dash line in Figure 2.1a. So the point when these two curves cross at left side corresponds to the smallest z_1 in this promising zone, within which the conditional power at this point is the minimal conditional power $CP_{\hat{\theta},min}$. This corresponds to $z_1 = 1.24$ and $CP_{\hat{\theta},min} = 0.3605$ in

Figure 2.1a. The upper bound of promising zone is the point when conditional power based on planned d_2 equals 0.9, which corresponds to $z_1 = 2.06$ and $CP_{\hat{\theta}}(z_1, d_2) = 0.9$. The black dotted and back medium-dash curves are the conditional powers based on original d_2 and adjusted d_2^* respectively; and both are against right y-axis in a scale ranging from 0 to 1 and coincide with each other outside the promising zone because d_2 is still used in these two zones. Conditional power based on adjusted d_2^* is boosted up in the range of $z_1 \in [1.24, 1.51]$ because the maximum allowable sample size $d_2^* = 135$ is used in the region due to the required number of events to gain power of 0.9 being larger than the maximum allowable limit; and be the constant of 0.9 for z_1 between 1.52 and 2.06. Figure 2.1b shows corresponding d_2^* with respect to z_1 in the promising zone.

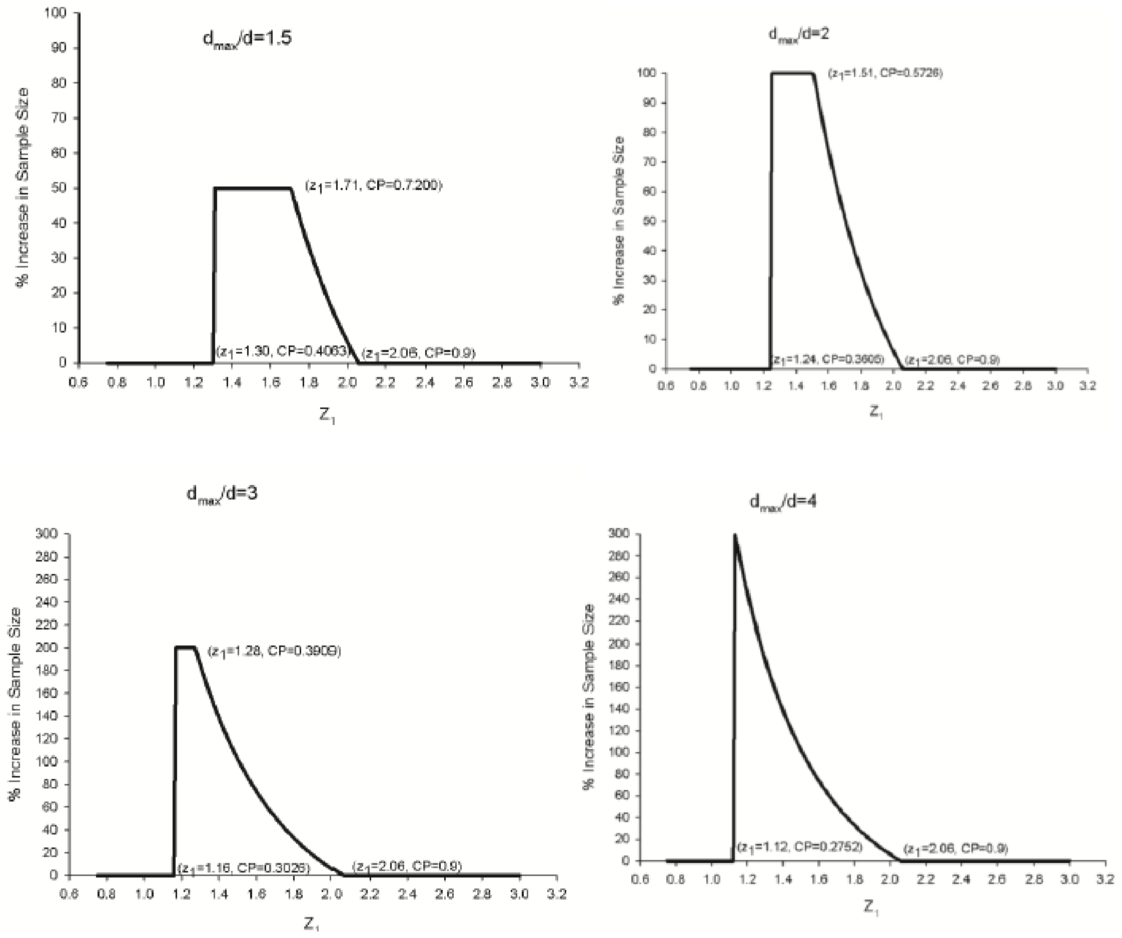


Figure 6(Fig. 2.2): Percent increase in Sample size

Figure 2.2: Percent increase in Sample size versus $z_1$for $\{t_1 = 0.5, 1\}$: Upper left for $\frac{d_{max}}{d} = 1.5$; upper right for Upper left for $\frac{d_{max}}{d} = 2$; lower left for $\frac{d_{max}}{d} = 3$; and lower right for $\frac{d_{max}}{d} = 4$

This promising zone is set up prior to trial start for a given set including $\alpha, \beta, \{t_1, 1\}$, d_{max}/d and a certain type of group sequential upper boundaries. $\alpha, \beta, \{t_1, 1\}$ and type of group sequential test defines $\{b_1, b\}$ and $\{d_1, d_2\}$ upfront. After conducting the trial to collect d_1 number of events, interim logrank test statistic z_1 will be calculated. If the conditional power $CP_{\hat{\theta}}(z_1, d_2)$ is located in the promising zone and null hypothesis is not rejected at interim, we continue into stage two to collect additional d_2^* (Figure 2.1) number of events such that $CP_{\hat{\theta}}(z_1, d_2^*)$=1- β if required number of events is below maximum allowable number or the same as the maximum allowable number when the required number exceeds it. When interim test statistic z_1 falls either the unfavorable zone or the favorable zone, the trial will continue to collect d_2 events with no adaptation.

Figure 2.2 shows the sample space division for z_1 when there is an equally spaced two-stage design with different ratios of maximum sample size after adaptation relative to pre-planned sample size. When $\frac{d_{max}}{d} = 1.5$, allowing maximum of 50% in total sample size increase, the promising zone starts from conditional power of 0.4063 to 0.9, corresponding to z_1 from 1.31 to 2.06. For $\frac{\mathrm{d_{max}}}{\mathrm{d}} = 2$, $\frac{\mathrm{d_{max}}}{\mathrm{d}} = 3$ and $\frac{\mathrm{d_{max}}}{\mathrm{d}} = 4$, the lower limit of promising zone is respectively with conditional power of 0.3605, 0.3026 and 0.2752. The lower bound of promising zone decreases as ratio of $\mathrm{d_{max}}/\mathrm{d}$ increases.

Section 2.4: CIBS I and II: Revisit

In old bad days, it took ten years from a failed, underpowered trial to a success trial conducted with enough power to detect alternative hypothesis. The estimated annual mortality rate from CIBS-II is 8.8% in the bisoprolol group and 13.2% in the placebo group. So the hazard ratio is estimated to be 1.5 (i.e., 0.132/0.088). Based on hazard ratio 1.5, for a two-stage group sequential trial with one-sided error of 0.025 and information vector of t = (0.5, 1) using WT boundary with shape parameter 0.15, the upper boundary vector is b_1= 2.554 and b=2.006. Total number of events required to detect hazard ratio 1.5 with above two-stage WT group sequential design is 261 (note that CIBS-II had 384 events in total in the end and CIBS-I only accumulated 120 events in total) when $\alpha = 0.025$ and $1 - \beta = 0.9$. If d_{max}/d is 3, a promising zone for CIBS-I can be constructed accordingly per steps in Section 2.3. Now let one take a look and see what would have been obtained had there been a sample increase implemented in CIBS-I while back to 1989? The minimal conditional power is then 0.3023 with optimal zone located within (0.3023, 0.9). From CIBS-I publication, interim log-rank test statistic was only 1.23 with low conditional power of 0.3531. Implementing optimal zone algorithm for survival data, additional 80 or more events are in need to be accumulated, rather than stopped the trial at the time when only 120 events were accumulated to disclaim the ‘failure’ of the trial. Had optimal zone method have been implemented, drug development time for bisoprolol would have been shorten up to maybe only 4-5 years instead of ten-year long plus huge economic cost for initiating one more trial.

Section 2.5: Simulation Results

Extensive simulations for proposed method are done with survival data in the presence or absence of censoring. As in Section 2.4, a one-sided two-stage group sequential design (GSD) is set up with WT boundaries with a shape parameter of 0.15 so that the upper bounds are defined accordingly: b=(2.554, 2.006) for equally spaced design with t = (0.5,1), b = (3.422, 1.963) for a design with early interim analysis (i.e., t = (0.2,1)); and b = (2.209, 2.043) with late interim analysis (i.e., t = (0.8,1)). Power requirement of 0.9 ($\beta = 0.1$) is used to search for total number of events to ensure enough power of detecting alternative hypothesis of hazard ratio of 2 (i.e., $\Delta = \frac{\lambda_c}{\lambda_E} = 2$). Subsequently, 90 total events is required for both equally spaced and late interim analysis, while only 88 is required for design with early interim analysis of t = (0.2, 1). Only exponential censoring with $\emptyset_C = \emptyset_E = 0.5\lambda_C$ is covered. That is: hazard rates of censoring in both treatment and placebo groups are the same and is 50% of the event hazard rate for placebo group subjects. Of course, censoring is assumed to be independent of both the time to event process and the accrual process. No futility boundaries are defined for simplicity but can be easily added if necessary. GSD is converted into adaptive GSD (A-GSD) by inserting an option of sample size increase in the situation when the interim result falls into the promising zone. To assess how an underpowered GSD performs under A-GSD, simulations are done with hazard ratio being 1.2, 1.4, 1.6, 1.8, and 2, in combination of different information vectors and $\frac{d_{max}}{d}$ ratios. In the meantime, the impacts of censoring on trial operating characteristics are shown as side results in both GSD and A-GSD.

Tables 2.1 – 2.4 list simulate operating characteristics with summaries of conditional results (Columns 5-7) and unconditional results (Columns 8-9) with Columns 5-7 being subset into two small columns with GSD and A-GSD side-by-side to illustrate resulting differences in between. Column 3 is the frequency distribution of three zones accompanied by Column 4 with probability of rejecting null hypothesis at interim given interim results, from which no rejection is present in both unfavorable and promising zones and only a portion of z_1 resulting in the right tail of the favorable zone have null hypothesis rejected at interim analysis. Columns 5 and 6 contain the conditional probability of rejecting null at final analysis and the combined conditional probability of rejecting null either at interim or final, respectively, conditional upon interim outcome. From Columns 5 and 6, it is shown that there is an obvious boost in conditional power after sample size adaptation when interim test statistic falling into the promising zone. Subsequently, Column 7 presents conditional average sample size per zone. Column 8, on the other hand, illustrates overall probabilities of rejection null at interim/final/interim or final and the expected average sample number irrespective of interim zone, following by the last column to show expected sample number for both GSD and A-GSD. As pointed out by a reviewer, conditional power is as important as overall power as the decision on any adaptation is taken at the time of the interim analysis and is therefore driven by the gain in conditional power and subsequently leading to increase in overall power.

Tables 2.1- 2.2 show the operating characteristics of both GSD and A-GSD for $\frac{d_{max}}{d} = 2$ with interim performed in the middle of the trial in the absence of censoring (Table 2.1) and in the presence of censoring (Table 2.2). In Table 2.1, the overall probability of rejecting null hypothesis under hazard ratio of 2 is 89.1% for GSD and increases to 92.0% with insertion of sample size increase in the promising zone. The increase in overall power from GSD to A-GSD is the largest when hazard ratio being 1.4 and 1.6. For example, it is 4.9% for $\Delta = 1.4$ (from 34.1% to 39.0%), 6.6% for $\Delta = 1.6$ (from 58.2% to 64.8%), 4.6% for $\Delta = 1.8$ (from 77.4% to 82.0%), and 2.9% for $\Delta = 2.0$ (from 89.1% to 92.0%). The increase of overall power using A-GSD is due to increase in conditional power when interim log-rank test statistic belongs to the promising zone. For instance, it is 17.5% for $\Delta = 1.4$ (from 45.5% to 63.0%) and 20.3% for $\Delta = 1.6$ (from 62.4% to 82.7%) and 15.1% for $\Delta = 1.8$ (from 77.9% to 93.0%). The designed parameters are calibrated at alternative hypothesis with hazard ratio of 2.0 and trial will be under-powered when the true hazard ratio is below 2.0. In this case, one can see that the proposed procedure rescued under-powered study to achieve a reasonable power (>=64.8%) as long as true hazard ratio is above 1.6. The increase in overall power is due to a considerable amount of patients falling in the promising zone (18.2%, 28.0%, 32.8%, 30.2% and 27.0% for Δ=1.2, 1.4, 1.6, 1.8 and 2.0 respectively). Table 2.1 has the same design as the one in Figure 2.1 as well as the one in the upper right corner in Figure 2.2. As depicted in Figure 2.1a, promising zone is an interval with z_1 ranging from 1.24 to 2.06, among which the maximum conditional power is 0.9 while first half (i.e., $z_1 \in [1.24,1.51]$) being less than 0.9. A boost in conditional power in the promising zone results in a boost in overall power, while the extent of increase decreases when true hazard ratio approaches the designed value of 2 because original group sequential design without sample size re-estimation already has large enough overall power. The average sample number (ASN) in Table 2.1 is consistently around 110 for A-GSD when the true hazard ratio is between 1.4 and 2.0, with 20+% increases from that of GSD.

From Table 2.2, there are no signs that inserting competing process of censoring will lower down overall power in either GSD or A-GSD as compared with cases in the absence of censoring. It seems that, uniformly for cases of $\Delta = 1.4$, 1.6, 1.8 and 2.0, powers in the presence of censoring are similar to their counterparts in Table 2.1. For hazard ratio 1.6, the overall powers are 58.4% and 64.9% for GSD and A-GSD respectively in the presence of censoring in Table 2.2 as compared with 58.2% and 64.8% in Table 2.1. Similarly, when true hazard ratio is 2, they are 89.1% and 92.0% in overall power for GSD and A-GSD respectively for cases in the absence of censoring in Table 2.1 as compared with 89.4% (for GSD) and 92.3% (for A-GSD) in the presence of censoring in Table 2.2.

Tables 2.3 – 2.4 investigate how the operating characteristics change if allowable increase ratio (i.e.,$\frac{d_{max}}{d}$) is changed from 2 to 4 in the absence of censoring (Table 2.3) and in the presence of censoring (Table 2.4) can rescue the underpowered trials better? And in what magnitude as compared with its respective cases in Tables 2.1 – 2.2? From Tables 2.3 - 2.4, increase in allowable sample size limit can increase overall power but in a small extent (from 92.0% to 92.4% in the absence of censoring in Tables 2.1 and 2.3 and from 92.3% to 93.0% in the presence of censoring in Tables 2.2 and 2.4 for $\Delta = 2.0$), but with expense of 13% percent in increase of expected sample size (113/127 to 111/126). Change in $\frac{d_{max}}{d}$ from 2 to 4 does not impact operation characteristics in all aspects except for impacts on expected sample size, which bring a question on the necessity of gaining that extra little power but at the expense of 13% of increase in sample size. Similarly for conditional power, for interim test statistic falling in the promising zone, which is the zone one wants to conduct rescue, conditional power increases up to 97.2% (vs. 96.6%) and 97.6% (vs. 97.0%) in the absence of censoring and in the presence of censoring respectively under $\frac{d_{max}}{d} = 4$ (vs. $\frac{d_{max}}{d} = 2$) at $\Delta = 2.0$.

Comparing with Tables 2.1 – 2.4, with which interims are done in the middle of the trial per pre-planned information level, t = (0.8,1) and t = (0.2,1) show the properties of A-GSD when the interim analysis performs in the later part and close to the end of the trial and at the early part of the trial, respectively. Power simulations to check impacts of timing design operation characteristics are not shown here. $t_1 = 0.2$ results in much less subjects falling in the promising zone while $t_1 = 0.8$ on the contrary results in more than half of first stage log-rank test statistic falling in the promising zone.

To assess rejecting probability under null hypothesis (i.e.,$\Delta = 1$), Table 2.5 presents operational characteristics of four scenarios in Tables 2.1 – 2.4. With no surprise, under null hypothesis the majority of subjects ended up in the unfavorable zone during simulations: 88.4% and 87.8% in the absence and presence of censoring respectively when $\frac{d_{max}}{d} = 2$ (vs. 89.4% and 87.3%

when $\frac{d_{max}}{d} = 4$). All simulations are done in 10000 simulation runs, type I error rates are all well-controlled as: 2.6%-2.9% for GSD with no sample size adaptation and 2.5%-2.7% for A-GSD with sample size increase when interim statistic falls in the optimal zone (Table 2.5).

Table 5(Tab. 2.1): Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design without censoring

Table 2.1: Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design without censoring while with t=(0.5,1), dmax/d=2, WT boundaries with shape parameter of 0.15, $\Delta = 2.0, \alpha = 0.025$ and $\beta = 0.1$.

Hazard Ratio for simulation	Interim outcome	P(interim outcome)	P(Rejection at interim | interim outcome)	P(Rejection at final | interim outcome)		Rejection Probability (interim or final) Conditional on Interim Outcome		E(d |Interim Outcome)		Overall Rejection Probability		E(d)	
				GSD	A-GSD	GSD	A-GSD	GSD	A-GSD	GSD (interim/final/either)	A-GSD (interim/final/either)	GSD	A-GSD
1.2	Unfavorable	74.4%	0%	4.0%	4.0%	4.0%	4.0%	90	90	2.8%/9.9%/12.7%	2.8%/11.2%/14.1%	90	101
	Promising	18.2%	0%	25.1%	32.3%	25.1%	32.3%	90	150				
	Favorable	7.4%	37.9%	32.5%	32.5%	70.4%	70.4%	90	90				
1.4	Unfavorable	55.0%	0%	13.0%	13.0%	13.0%	13.0%	90	90	7.3%/26.8%/34.1%	7.3%/31.7%/39.0%	90	107
	Promising	28.0%	0%	45.5%	63.0%	45.5%	63.0%	90	148				
	Favorable	17.0%	43.0%	40.6%	40.6%	83.6%	83.6%	90	90				
1.6	Unfavorable	37.2%	0%	26.8%	26.8%	26.8%	26.8%	90	90	15.7%/42.5%/58.2%	15.7%/49.2%/64.8%	90	111
	Promising	32.8%	0%	62.4%	82.7%	62.4%	82.7%	90					
	Favorable	30.0%	52.2%	40.3%	40.3%	92.5%	92.5%	90	90				
1.8	Unfavorable	24.1%	0%	41.3%	41.3%	41.3%	41.3%	90	90	27.2%/50.2%/77.4%	27.2%/54.8%/82.0%	90	112
	Promising	30.2%	0%	77.9%	93.0%	77.9%	93.0%	90	142				
	Favorable	45.7%	59.6%	36.6%	36.6%	96.2%	96.2%	90	90				
2.0	Unfavorable	15.4%	0%	58.9%	58.9%	58.9%	58.9%	90	90	38.6%/50.5%/89.1%	38.6%/53.4%/92.0%	90	113
	Promising	27.0%	0%	86.0%	96.6%	86.0%	96.6%	90	139				
	Favorable	57.6%	67.0%	31.7%	31.7%	98.7%	98.7%	90	90				

Table 6(Tab. 2.2): Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design in the presence of censoring

Table 2.2: Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design in the presence of censoring while with t=(0.5,1),dmax/d=2, WT boundaries with shape parameter of 0.15, $\Delta = 2.0, \alpha = 0.025$ and $\beta = 0.1$.

Hazard Ratio for simulation	Interim outcome	P(interim outcome)	P(Rejection at interim | interim outcome)	P(Rejection at final | interim outcome)		Rejection Probability (interim or final) Conditional on Interim Outcome		E(d |Interim Outcome)		Overall Rejection Probability		E(d)	
				GSD	A-GSD	GSD	A-GSD	GSD	A-GSD	GSD (interim/final/either)	A-GSD (interim/final/either)	GSD	A-GSD
1.2	Unfavorable	73.9%	0%	3.9%	3.9%	3.9%	3.9%	90	90	2.5%/10.2%/12.6%	2.5%/11.8%/14.3%	90	102
	Promising	18.9%	0%	26.0%	34.9%	26.0%	34.9%	90	150				
	Favorable	7.3%	34.3%	32.6%	32.6%	66.9%	66.9%	90	90				
1.4	Unfavorable	55.5%	0%	12.9%	12.9%	12.9%	12.9%	90	90	7.3%/26.7%/34.0%	7.3%/32.2%/39.5%	90	107
	Promising	27.5%	0%	46.1%	66.0%	46.1%	66.0%	90	147				
	Favorable	16.9%	42.8%	40.6%	40.6%	83.4%	83.4%	90	90				
1.6	Unfavorable	37.6%	0%	27.0%	27.0%	27.0%	27.0%	90	90	15.6%/42.9%/58.4%	15.6%/49.3%/64.9%	90	111
	Promising	32.0%	0%	64.0%	84.0%	64.0%	84.0%	90	144				
	Favorable	30.4%	51.3%	40.2%	40.2%	91.5%	91.5%	90	90				
1.8	Unfavorable	23.6%	0%	41.4%	41.4%	41.4%	41.4%	90	90	26.2%/51.5%/77.7%	26.2%/56.3%/82.6%	90	112
	Promising	31.7%	0%	77.8%	93.1%	77.8%	93.1%	90	141				
	Favorable	44.8%	58.5%	38.3%	38.3%	96.8%	96.8%	90	90				
2.0	Unfavorable	14.5%	0%	58.0%	58.0%	58.0%	58.0%	90	90	38.5%/50.9%/89.4%	38.5%/53.8%/92.3%	90	111
	Promising	27.5%	0%	86.4%	97.0%	86.4%	97.0%	90	138				
	Favorable	58.0%	66.3%	32.3%	32.3%	98.6%	98.6%	90	90				

Table 7(Tab. 2.3): Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design without censoring

Table 2.3: Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design without censoring while with t=(0.5,1), dmax/d=4, WT boundaries with shape parameter of 0.15, $\Delta = 2.0, \alpha = 0.025$ and $\beta = 0.1$.

Hazard Ratio for simulation	Interim outcome	P(interim outcome)	P(Rejection at interim | interim outcome)	P(Rejection at final | interim outcome)		Rejection Probability (interim or final) Conditional on Interim Outcome		E(d |Interim Outcome)		Overall Rejection Probability		E(d)	
				GSD	A-GSD	GSD	A-GSD	GSD	A-GSD	GSD (interim/final/either)	A-GSD (interim/final/either)	GSD	A-GSD
1.2	Unfavorable	69.0%	0%	3.5%	3.5%	3.5%	3.5%	90	90	2.9%/10.6%/13.4%	2.9%/13.9%/16.8%	90	116
	Promising	23.2%	0%	23.5%	38.0%	23.5%	38.0%	90	197				
	Favorable	7.8%	36.9%	34.1%	34.1%	70.9%	70.9%	90	90				
1.4	Unfavorable	50.3%	0%	11.5%	11.5%	11.5%	11.5%	90	90	7.6%/26.9%/34.5%	7.6%/35.5%/43.1%	90	124
	Promising	32.3%	0%	44.3%	71.0%	44.3%	71.0%	90	187				
	Favorable	17.4%	43.5%	38.9%	38.9%	82.4%	82.4%	90	90				
1.6	Unfavorable	32.6%	0%	24.0%	24.0%	24.0%	24.0%	90	90	16.1%/42.2%/58.3%	16.1%/52.2%/68.2%	90	129
	Promising	36.6%	0%	60.0%	87.1%	60.0%	87.1%	90	180				
	Favorable	30.9%	52.0%	40.5%	40.5%	92.4%	92.4%	90	90				
1.8	Unfavorable	21.2%	0%	40.5%	40.5%	40.5%	40.5%	90	90	26.7%/51.0%/77.7%	26.7%/57.4%/84.1%	90	127
	Promising	34.2%	0%	75.9%	94.6%	75.9%	94.6%	90					
	Favorable	44.7%	59.7%	37.0%	37.0%	96.7%	96.7%	90	90				
2.0	Unfavorable	12.3%	0%	52.2%	52.2%	52.2%	52.2%	90	90	38.3%/50.4%/88.7%	38.3%/54.1%/92.4%	90	127
	Promising	29.3%	0%	84.6%	97.2%	84.6%	97.2%	90	169				
	Favorable	58.4%	65.5%	32.9%	32.9%	98.4%	98.4%	90	90				

Table 8(Tab. 2.4): Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design in the presence of censoring

Table 2.4: Simulated Operating Characteristics of Adaptive or Non-adaptive Group Sequential Design in the presence of censoring while with t=(0.5,1),dmax/d=4, WT boundaries with shape parameter of 0.15, $\Delta = 2.0, \alpha = 0.025$ and $\beta = 0.1$.

Hazard Ratio for simulation	Interim outcome	P(interim outcome)	P(Rejection at interim | interim outcome)	P(Rejection at final | interim outcome)		Rejection Probability (interim or final) Conditional on Interim Outcome		E(d |Interim Outcome)		Overall Rejection Probability		E(d)	
				GSD	A-GSD	GSD	A-GSD	GSD	A-GSD	GSD (interim/final/either)	A-GSD (interim/final/either)	GSD	A-GSD
1.2	Unfavorable	69.9%	0%	3.4%	3.4%	3.4%	3.4%	90	90	2.5%/10.2%/12.6%	2.5%/13.6%/16.1%	90	115
	Promising	22.9%	0%	23.7%	38.9%	23.7%	38.9%	90	196				
	Favorable	7.3%	34.3%	32.6%	32.6%	66.9%	66.9%	90	90				
1.4	Unfavorable	50.8%	0%	11.2%	11.2%	11.2%	11.2%	90	90	7.3%/26.7%/34.0%	7.3%/35.5%/42.7%	90	123
	Promising	32.3%	0%	43.8%	70.9%	43.8%	70.9%	90	186				
	Favorable	16.9%	42.8%	40.6%	40.6%	83.4%	83.4%	90	90				
1.6	Unfavorable	32.9%	0%	24.0%	24.0%	24.0%	24.0%	90	90	15.6%/42.9%/58.4%	15.6%/52.5%/68.0%	90	130
	Promising	36.8%	0%	61.8%	88.0%	61.8%	88.0%	90	181				
	Favorable	30.4%	51.3%	40.2%	40.2%	91.5%	91.5%	90	90				
1.8	Unfavorable	20.0%	0%	38.3%	38.3%	38.3%	38.3%	90	90	26.2%/51.5%/77.7%	26.2%/58.1%/84.3%	90	129
	Promising	35.2%	0%	75.8%	94.4%	75.8%	94.4%	90	173				
	Favorable	44.8%	58.5%	38.3%	38.3%	96.8%	96.8%	90	90				
2.0	Unfavorable	12.0%	0%	54.5%	54.5%	54.5%	54.5%	90	90	38.5%/50.9%/89.4%	38.5%/54.5%/93.0%	90	126
	Promising	29.9%	0%	85.5%	97.6%	85.5%	97.6%	90	165				
	Favorable	58.0%	66.3%	32.3%	32.3%	98.6%	98.6%	90	90				

Table 9(Tab. 2.5): Simulated Type I error for eight different designs

Table 2.5: Simulated Type I error for eight different designs which have WT boundaries with shape parameter of 0.15, $\Delta = 2.0, \alpha = 0.025$ and $\beta = 0.1$.

	Interim outcome	P(interim outcome)	Overall Rejection Probability		E(d)	
			GSD (interim/final/either)	A_GSD (interim/final/either)	GSD	A-GSD
$\Delta = 1.0$ (simulation), $\varphi = 0$, t = (0.5,1), $d_{max}/d = 2$	Unfavorable	88.4%	0.7%/2.2%/2.9%	0.7%/2.0%2.7%	90	96
	Promising	9.2%				
	Favorable	2.4%				
$\Delta = 1.0$ (simulation),$\varphi = 0.5\lambda_c$, t = (0.5,1), $d_{max}/d = 2$	Unfavorable	89.4%	0.5%/2.1%/2.6%	0.5%/2.0%/2.5%	90	96
	Promising	8.7%				
	Favorable	1.9%				
$\Delta = 1.0$ (simulation), $\varphi = 0$, t = (0.5,1), d_{max}/d =4	Unfavorable	87.8%	0.6%/2.2%/2.8%	0.6%/2.0%/2.6%	90	102
	Promising	10.4%				
	Favorable	2.3%				
$\Delta = 1.0$ (simulation),$\varphi = 0.5\lambda_c$, t = (0.5,1), $d_{max}/d = 4$	Unfavorable	87.3%	0.5%/2.1%/2.6%	0.5%/2.0%/2.5%	90	102
	Promising	10.8%				
	Favorable	1.9%				

Section 2.6: Discussion

This paper extends Mehta and Pocock (2010) to survival trials with a real example from a historical drug development example, together with extensive simulations on various scenarios in the presence or absence of censoring, large or moderate allowable limit in sample size increase, interim analysis occurring at an earlier or later time point. It can be seen that this method is very easy to implement for survival data and can be presented to non-statisticians easier than other methods as conventional test statistic and original critical value will be used for final analysis, which hence avoids the hotly-debated issue of violating “one patient one vote” with weighted test for final analysis. Due to the fact that no real clinical trials are lack of censoring, which can be caused by early withdrawal due to adverse event, lack of efficacy, loss to follow-up and subject consent as well as administratively censoring at analysis time point, simulation results for cases in the presence of censoring will assure its practicability in survival group sequential trials. Adaptation method proposed here performs well when timing of interim is not so early. Doing adaptation too early should not be considered in general as estimate of drug effect is not stable at the early stage, thus downgrading the capability of rescuing an underpowered trial by sample size increase in the promising zone. Results also show that after a certain level, further increase of allowable sample size limit will barely help in terms of conditional and overall powers but at a big expense of expected sample size, therefore economically not efficient for having $\frac{d_{max}}{d}$ too large.

In the past two decades, numerous publications on sample size re-estimation and adaptive designs are mainly from two aspects: 1)use weighted test to construct a final test statistic comparing with original critical value, with which weighted test has the same distributional property under null hypothesis as the planned test statistic so that the type I error rate is controlled; 2) use conventional test even after adaptation but adjust critical value so that the overall type I error rate is controlled when decision is based on using conventional test statistic to be compared with adjusted critical value. Sample size increase in the promising zone provided the third way to control type I error rate. That is to define promising zone upfront based on type I error, power, budget limit, data type and test statistic to be used for both interim and final analyses together with adaptation rules in the promising zone as in Figures 2.1a and 2.1b. In this promising zone, sample size can be increased and conventional test without weighting strategy will be used to compare with the original critical value without any adjustment. Although being quite novel, this is a method not well-evaluated yet. As being criticized by Emerson, Levin and Emerson (2011), the efficiency of this method under-investigated. Therefore, how to improve the efficiency of this method in terms of minimizing average sample size with respect to parameters of interests is the direction for future research. All in all, the promising zone is defined as the region of z_1 (or equivalently the region of conditional power under the initial design) where$b^*(z_1, d_2^*) \leq b$. The motivation for defining the promising zone in this way is that one can use the regular test $Z^* \geq b$ for the final analysis without scarifying type I error rate control. However, as pointed out by the reviewer and agreed by the authors, that this is by no means the only way to specify the promising zone. In general, the promising zone could simply be perceived as a region of z_1 within which the sponsor is willing to increase the sample size in exchange for a substantial gain in conditional power. It may be convenient to confine it to a region within which the conventional test $Z^* \geq b$ is valid, but this is not necessary. If the promising zone contains a region in which $b^*(z_1, d_2^*) > b$, one would control the type I error rate with the CHW test $Z_{CHW}^* \geq b$. The choice of promising zone and the method for controlling the type I error is not necessarily linked.

Reference

Bauer, P. and Kohne, K. (1994), Evaluation of Experiment with Adaptive Interim Analysis. *Biometrics*, 50:1029-1041.

CIBIS-I: CIBIS Investigators and Committees .A randomized trial of beta-blockade in heart failure: The Cardiac Insufficiency Bisoprolol Study (CIBIS) (1994). Circulation. 1994 Oct;90(4):1765-73.

CIBIS-II: CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol II (CIBIS-II): a randomised trial (1999). *Lancet*, 353: 9-13.

Chen, Y.H., DeMets, D.L, Lan, K.K. (2004). Increase the sample size when the unblinded interim results is promising. *Stat Med*, 23:1023-38.

Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B* 1972; 34(2): 187–220.

Cui, L., Hung, H.M., Wang, S.J. (1999), Modification of sample size in group sequential clinical trials. *Biometrics,* 55:853-7.

Emerson, S., Levin, G.P., and Emerson, S.C. (2011), Comments on "Adaptive Increase in sample size when interim results are promising: A practical guide with examples". *Stat Med*, 00:1-16.

FDA Guidance for industry: Adaptive design clinical trials for drugs and biologics (2010).

Gao, P., Ware, J.H, Mehta, C.R (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *J Biopharm Stat*, 18: 1184-96.

Jennison, C and Turnbull, B.W. (2000), Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman and Hall.

Mehta, C.R., Pocock, S.J. (2011). Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples. *Stat Med,* 30(28):3267-84.

Muller, H.H. and Schafer, H. (2001), Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57:886-91.

Proschan, M.A. and Hunsberger, S.A. (1995), Designed Extension of Studies Based on Conditional Power. *Biometrics*, 51:1315-1324.

PhRMA White paper of the ohrma adaptive working group (2007). *DIA Journal*.

Wassmer, G. (1998), A Comparison of Two Methods for Adaptive Interim Analysis in Clinical Trials. *Biometrics,* 54:696-705.

Proschan, M., Lan, K., and Wittes, J. (2006). Statistical Monitoring of Clinical Trials: A Unified Approach. Springer, New York.17: Wang, S.K. and Tsiatis A.A. (1987). Approximatelyoptimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193-199.

Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal,* 48 (4), 714-729.

CHAPTER 3

Prediction of the Timing of Events in Clinical Trials with Survival Endpoints: A Trial Example

(Being Reviewed by *Journal of Biopharmaceutical Statistics*)

Abstract: In event-based clinical trials, interim and final analyses at pre-specified event times are often proposed. In a randomized withdrawal trial with time-to-event primary endpoint, the design consists of subjects receiving a test treatment for a specified period and then being randomized to continue on that treatment or placebo. We present methodology to predict the time of reaching a required number of events during the double-blind phase of such a trial. We consider prediction at any time during the course of this trial: at the beginning of the trial; during the open-label phase of the trial and also during the double-blind phase of the trial (where some subjects could still be in the open-label phase). There has been recent work on tackling various aspects of this problem using parametric, semi-parametric or from a Bayesian perspective. Starting from Whitehead's method (2001), we consider four additional features: (i) censoring process can be incorporated; (ii) calculating expected number of events by a future calendar time, t_2, for subjects who were in the risk set at t_1; (iii) predicting number of events by a future time point t_2 for subjects who were enrolled prior to randomization and will be randomized at a fixed time point before t_2; and (iv) various parametric survival distributions other than exponential (i.e., Weibull, Lognormal, Log logistic). We applied our methodology during the conduct of a recently completed clinical trial to accurately predict the timing of the interim analysis. This allowed sufficient resources to be deployed leading to timely data analysis and reporting.

Keywords: Time-to-event outcomes; trial duration prediction; interim analysis; survival endpoint.

Section 3.1: Introduction

In clinical trials designed to compare survival curves under two treatments, it is often desirable to model and predict the timing to a pre-specified number of events since this has important implications on resource allocation, study budget and planning. In a randomized withdrawal trial with time-to-event primary endpoint, the design consist of subjects receiving a test treatment for a specified period of time (herein referred to as open-label phase) and then being randomly assigned to continue on that treatment or placebo (herein referred to as double-blind phase). During the recruitment period, subjects who meet inclusion and exclusion criteria are screened, enter the trial at a certain rate for a specified period of time, and if meeting certain stability requirements, are randomized to one of the two treatment groups. After randomization process ends, there is a period called “continuation period”, during which patient follow-up continues (on treatment or placebo). Aside from having events during the trial, some event times (e.g., death or relapse times) are typically not observed and are said to be right-censored, as death times are only known to be greater or equal to the censoring time. Two types of censorship exist: 1) Subjects withdraw early due to adverse events, withdrawal of consent or loss to contact. These censorings are generally called “loss to follow-up”; 2) Subjects remain event-free at time of study termination, and are said to be “administratively censored”. Both censorships are not related to individual death times; hence it seems reasonable to assume independence between event and censoring time in prediction and statistical analysis. The log-rank statistic (Mantel, 1966) has been widely accepted and used to compare survival curves in the presence of such censorships. Simulations (Lee, Desu and Gehan, 1975) show that the Mantel statistic (logrank) has acceptable power against other types of alternatives as well as proportional hazards, in which one hazard is a constant multiple of the other.

In the literature, some authors have considered the dual problem of planning the size (i.e., the required number of patients) and the required duration of the trial when death times are assumed to be exponentially distributed. Pasternack and Gilbert (1971) converted fixed sample size determination into equivalent “person-years at risk”. When patients were accrued by cohorts, they derived required duration and number of events to ensure enough power to detect a certain percentage increase in the median survival of subjects in treatment group over the control group. Similar to Pasternack and Gilbert (1971), George and Desu (1974) also assumed exponential death times in the situation with lack of censoring during the trial. Instead of accrual by cohorts, accumulated number of patient-year in the time interval to obtain required number of events is now modeled as a filtered Poisson process. George and Desu (1974) showed that the required duration can be found by solving a non-linear equation using iterative techniques and proved that the minimal (optimal) required duration of study requires no continuation period after accrual period. Rubinstein, Gail and Santner (1981) extended the trial length calculations of Pasternack and Gilbert (1971) and George and Desu (1974) to cover experiments with Poisson accrual, loss to follow-up and a continuation period. In the case of no loss to follow-up and no continuation period, Rubinstein, Gail and Santner's (1981) calculations differ very little from Table 2 of George and Desu (1974). All these length calculations are based on the assumption that the death times are exponential and the comparison was made via the maximum-likelihood-estimation (MLE) of the death hazard rates. Simulations in Rubinstein, Gail and Santner (1981) showed that trial length calculations using MLE yield accurate power for Logrank test for exponential death times and approximately valid even for Weibull death times. Although we use death times and survival time interchangeably, survival endpoints have actually become more broadly used, including not only time to death endpoint, but also time to other events. In a randomized withdrawal study design, subjects receiving a test treatment (i.e., open-label) for a fixed-period of time are randomly assigned to continue on test drug or switch to placebo (i.e., withdrawal of active therapy) in the double-blind phase. Any difference that appears between the group continuing on test drug and the group randomized to placebo would demonstrate the effect of the active treatment. For example, in randomized withdrawal trials, time from randomization to relapse in the double-blind phase is the key efficacy variable (measuring persistence of effectiveness) used to compare treatments in the double-blind phase after subjects being stabilized for disease symptoms in the open-label phase. See more details on randomized withdrawal trials on Pages 17-19 of FDA guidance document "Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products". Interim analyses are a common feature of clinical trials, especially for large trials or trials for rare disease with a low accrual rate. Whitehead (2001) described predicting final sample size and total duration of a sequential survival study with exponential death times and examined the pay-off between speed of accrual rate and length of continuation period, however, competing process of time to follow-up (censoring) was not considered. All prediction methods first estimate number of events required to test null hypothesis with certain power and size level, and then length of trial is estimated using accrual rate, rate of time to event, rate of loss to follow-up, accrual time, and length of the continuation period. In this paper, we extend Whitehead (2001) to include censoring process in prediction prior to trial start and then provide methods to carry out prediction during the trial prior to interim analysis.

In addition to parametric and semi-parametric approaches, others have considered prediction algorithms using Bayesian methods. Posterior parameters can be sampled using Markov Chain Monte Carlo (MCMC) with help from priors, observed likelihood at time of prediction assuming parametric exponential survival times (Bagiella and Heitjan, 2001) or Weibull survival times (Ying and Heitjian, 2008). The predictive probability distribution of calendar time to obtain certain number of events can be completed by simulation based on posterior parameters for subjects not yet having an event at prediction time or to be recruited with a homogenous accrual process. Cumulative events at future time t_2 consist of events occurring prior to and after prediction time. When randomization is blinded, estimating of posterior probability of being in one particular treatment can be incorporated in the middle of sampling algorithm (Donovan, Elliott and Heitjan, 2006); and similar research was done in the situation when randomization is not only masked but also blocked (Donovan, Elliott and Heitjan, 2007). Additional variation includes prediction when there is a delay in reporting events during the trial with some withdrawals recorded in database possibly having had an event occurred prior to withdrawal but without reporting (Wang et al., 2012). All of these predictions assumed homogeneous accrual process together with either exponential or Weibull event times. Non-homogenous accrual combined with Bayesian prediction have also been explored in order to take into account different accrual rates across regions that normally occur in multi-regional clinical trials (Zhang and Long, 2010, 2012a). Additionally, Zhang and Long (2012b) published a systematic review paper on modeling and prediction of subject accrual and event times in clinical trials using Bayesian methods.

Although extensive research has been done for various situations from Bayesian perspective, the choice of prior distribution, extensive sampling for each posterior parameter and creating complete sample based on posterior parameters somehow prevent this set of methods from being widely used in clinical trials because of their computational and methodological complexities. Commercial software developers are now beginning to fill that need.

In this paper, we develop methodologies to carry out prediction during the trial with or without censoring using different parametric death time distributions. Use of accumulated trial data can be incorporated without unmasking study treatment.

Section 3.2: Statistical Methods: Set Up

To compare two treatments based on survival response, hypothesis testing could be constructed on a summary measure of the log hazard ratio, $\theta = -log\frac{\lambda_E(t)}{\lambda_C(t)}$, for all $t > 0$, where $\lambda_E(t)$ and $\lambda_C(t)$ denote hazard at time t for experimental and control group respectively, when there is exponential death time or the more general case of constant hazard ratio over time. The null hypothesis is $H_0: \theta = 0$ against $H_A: \theta = \theta_R$, where θ_R is the clinically meaningful difference that the experimental group holds over the control group. Alternatively, this referential difference can be characterized in terms of survival functions $S_E(t)$ and $S_C(t)$ on E (experimental group) and C (control group): $\theta = -\log[-\log[S_E(t)\ ]\ ] + \log[-\log[S_C(t)\ ]\ ]$, for all $t > 0$. After finding survivor probability for control group $S_C(t_0)$ at t_0, a specified $S_E(t_0)$ can be estimated. If the probability of rejecting (required power) null hypothesis against alternative is $1-\beta$ at two-sided significance level α, utilizing asymptotical normality properties of Logrank statistic, the required number of events is $e = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\frac{\theta_R}{2})^2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quartile value of standard normal random variable. This is deduced from the fact that, when θ is small, the logrank statistic Z is approximately normal and distributed with mean θV and variance V, and $V = e/4$ (Section 9.2.1 of Collett, 1994), e is expected number of events at the end of the trial. Asymptotical normal approximation is very accurate for $\theta_R < 1$; and acceptable for $1 \leq \theta_R < 2$. We only make use of relative reference of θ_R, α and β. The rate of randomization accrual, randomization accrual time, length of continuation period, and rate of loss to follow-up haven't played a role in trial design at this stage. Starting from required number of events, number of patients to be randomized in time T and then to be followed in time τ can be deduced in Sections below. We specifically consider prediction based on a clinical trial with a randomized withdrawal design. We illustrate predicting number of patients to recruit (or trial duration to achieve required number of events) by two different scenarios: 1) Predicting before trial start. In this case, we can integrate with respect to distribution of times from entry to end of trial (Figure 3.1a, Appendix 3.1). 2) Predicting during the trial before interim or final analysis (Figure 3.1b, Appendices 3.2 and 3.3). During the trial, specifically at time t_1, the expected cumulative number of events up to a future time t_2 includes events that have occurred prior to and on t_1 plus events yet to occur between t_1 and t_2. For trials which have fixed-length phases prior to randomization into the double-blind phase, at time t_1, some subjects may be ongoing during the phases prior to randomization and will be randomized at a known time between t_1 and t_2, this cohort will contribute to the total number of events up to a future time t_2. Subjects who were randomized but remained event free in the double-blind phase of the trial at t_1, who are in the at-risk set at predicting time t_1, will also contribute to future events between t_1 and t_2. Figure 3.1a depicts the prediction prior to trial start and Figure 3.1b depicts prediction during the trial. Appendix 3.1 shows prediction algorithm in the presence of censoring prior to trial start. Appendix 3.2 describes prediction algorithm for subjects to be randomized at a known time between t_1 and t_2 with or without censoring with death times of exponential, Weibull, Log-logistic and Lognormal respectively and exponential censoring when it is present. Appendix 3.3 provides the prediction method for subjects who are in the risk set at prediction time t_1.

To do prediction prior to trial start, as depicted in Figure 3.1a, subjects are uniformly randomized in time interval $[0, T]$ months. After randomization period, subjects remained in the trial are continued to be followed for additional τ months. Time to event or time to loss to follow-up can occur at any time during period $[0, T + \tau]$. Subjects who are still remained event-free at time $T + \tau$ are administratively censored. Appendix 3.1 describes calculation of expected number of events by time $T + \tau$, provided that both survival and censoring times are exponentially distributed and there is an uniform randomization period. From Figure 3.1a, where there is approximate uniform randomization accrual in $[0, T]$ and subjects who have remained in the trial at time T are all followed for additional τ months. From Figure 3.1a, we have 9 events and 4 censoring by time $T + \tau$, including one with administrative censoring.

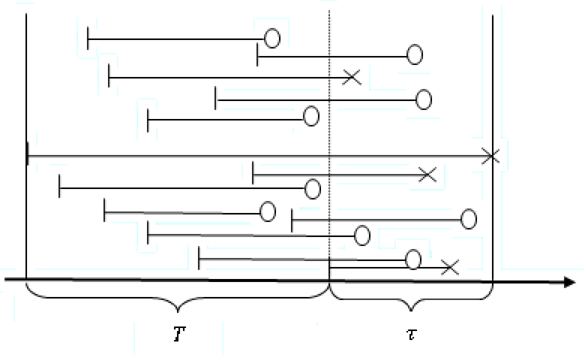


Figure 7(Fig. 3.1): Depiction of prediction prior to and during trial start

Figure 3.1: Depiction of prediction prior to and during trial start.
Figure 3.1a: Depiction of prediction prior to trial start with hypothetical subjects. Vertical bar "|" on the left hand of time line denotes the timing of performing randomization procedure and then subjects entered into the double-blind phase. Circle on the right hand indicates survival event occurred on this subject during the double-blind phase while cross symbol denotes censoring.

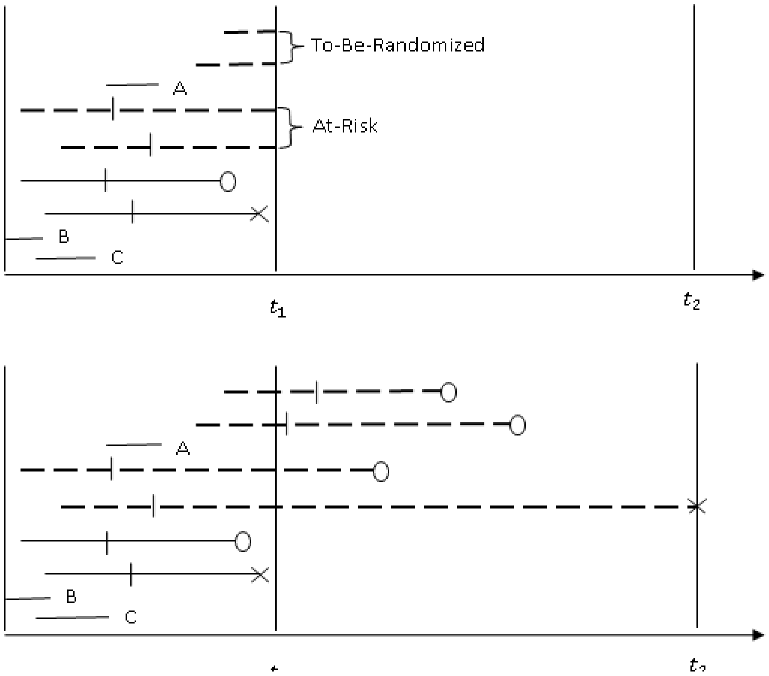


Figure 3.1b: Depiction of prediction during the trial. Upper graph: hypothetical subjects status before prediction at t_1. Lower graph: subjects status by time t_2. Vertical bar and circle symbols are defined in the same way as in Figure 3.1a.

Prediction is not a one-time thing and it is not just required prior to trial start. In normal practice, data can be blindly reviewed in order to obtain more information about what is going on in the trial while still not unblinding treatment information in order to maintain trial validity.

Different from Figure 3.1a, subjects in Figure 3.1b start the trial with a fixed-length period prior to randomization. For example, in a randomized withdrawal trial, subjects will be treated in an open-label period with study medication to stabilize acute symptoms before being randomized to continue on the study drug or being switched to placebo. Time from randomization to first documentation of relapse in the double-blind phase is primary endpoint to be observed so that the superiority of study drug over placebo in terms of delaying time to relapse can be assessed.

As illustrated in the upper half of Figure 3.1b, there were three subjects who withdrew early in the phases prior to randomization (Subjects A, B and C). At time t_1, one subject already had an event and one was censored; and four subjects who remained in the trial at time t_1, within which two out of four will be randomized between t_1 and t_2, the other two were randomized prior to t_1 and are considered to be in the at risk set and might have events in $(t_1, t_2]$. As illustrated in lower half of Figure 3.1b, by time t_2, the accumulated number of events in the double-blind phase can come from three different resources:

- events occurred prior to or on t_1
- from subjects who are in the phases prior to double-blind phase and will be randomized between $t_1$and t_2, who could have events by time t_2

- from subjects who are in the risk set at t_1, who may have events by time t_2

Starting from cases well-known in the literature, Section 3.3 first extends Whitehead (2001) to predict trial duration for newly randomized subjects in the presence of censoring, assuming time to censoring non-related to death time in the trial. Besides working out predicting trial duration prior to trial start in the presence of censoring while Whitehead (2001) only has the case without censoring (i.e. $\phi_c = \phi_E = 0$), our main contributions are mainly in Sections 3.4 and 3.5 for prediction during the trial in the absence or presence of censoring. As depicted in Figure 3.1b, subjects who are ongoing at prediction time t_1 consist with two cohorts: "To-Be-Randomized" subjects who are still ongoing in the phases prior to the double-blind phase and "At-Risk" subjects who are ongoing without events in the double-blind phase at time of prediction. Section 3.4 is about prediction for "To-be-randomized" subjects who will be randomized at a known time between t_1 and t_2, starting with the case with censoring (Section 3.4.1) to the case without censoring, and from exponential death times to non-exponential death times (Section 3.4.3). Section 3.5 describes prediction of expected number of events for "at-risk" subjects. Section 3.5.1 starts with the simpler case of no censoring present in the trial under exponential death time. Section 3.5.2 is for exponential death time in the presence of exponential censoring. Section 3.5.3 explores other death times in the presence of exponential censoring.

Section 3.3: Prediction Prior to Trial Start in the Presence of Censoring

In case that time to censoring competes with the time to event process in the double-blind phase, subjects can be censored prior to a particular calendar time. If censoring is indeed present in the trial, ignoring its existence will result in overestimating number of events at a given time and consequently underestimate the required trial duration needed to obtain certain number of events for analysis.

Now let's consider the prediction of trial duration prior to trial start. The steps to implement prediction for number of events prior to trial start or for newly randomized subjects are depicted in Appendix 3.1. Subjects are uniformly randomized at a rate of a for T months, resulting in aT subjects randomized over time interval $[0, T]$. Since randomization ratio is $A: 1$ for treatment group over control group, the expected number randomized into treatment and control group are $\frac{A}{A+1} aT$ and $\frac{1}{A+1} aT$, respectively. Because subjects are uniformly randomized into the double-blind phase over $[0, T]$, the times from being randomized to end-of-study (EOS) are also independent and identically distributed (i.i.d) uniform over [τ, T+ τ] with density $\frac{1}{T}$ (where τ is the follow-up time). Given a time interval u from entry onto control group to end-of-study, the probability that this entry will result in an event is $\frac{\lambda_C}{\lambda_C + \phi_C} [\, 1 - \exp[-(\lambda_C + \phi_C)u]\,]$ given that time to event is i.i.d. exponential (λ_C), time to censoring is i.i.d. exponential (ϕ_C); and the two processes are independent of each other. Based on uniform distribution of u (i.e. the times from entry to end-of-study (EOS)) and given n_C subjects being randomized into the control group, the expected number of events achieved by time T+ τ in the control group is:

$$E(e_c|n_C) = \frac{\lambda_C n_C}{T(\lambda_C + \phi_C)} \left[T + \frac{\exp[-(\lambda_C + \phi_C)(T+\tau)] - \exp[-(\lambda_C + \phi_C)\tau]}{\lambda_C + \phi_C} \right].$$

Replacing n_C with $E(n_C)$, we get the expected number of events in the control group for newly randomized subjects by time T+ τ as follows:

$$E(e_c) = \frac{a\,\lambda_C}{(A+1)\,(\lambda_C+\phi_C)}\left[T + \frac{\exp[-(\lambda_C+\phi_C)(T+\tau)]-\exp[-(\lambda_C+\phi_C)\tau]}{\lambda_C+\phi_C}\right].$$

The process of conditioning and un-conditioning are repeatedly used in above formulation and the conditional independence between death times and censoring times do play a key role in finding the probability of resulting in an event rather than being censored by a particular time. Treatment group follows the same procedure as the control group. Adding up events in both groups leads to the predicted number of events by T+ τ for newly randomized subjects in the presence of censoring. That is:

$$E(e) = E(e_C) + E(e_E) =$$

$$\frac{aT\,\lambda_C}{(A+1)\,(\lambda_C+\phi_C)} + \frac{aA\,T\lambda_E}{(A+1)\,(\lambda_E+\phi_E)} + \frac{a\,\lambda_C\,[\exp[-(\lambda_C+\phi_C)(T+\tau)]-\exp[-(\lambda_C+\phi_C)\tau]]}{(A+1)\,(\lambda_C+\phi_C)^2} +$$

$$\frac{aA\,\lambda_E[\exp[-(\lambda_E+\phi_E)(T+\tau)]-\exp[-(\lambda_E+\phi_E)\tau]]}{(A+1)\,(\lambda_E+\phi_E)^2}\ .$$

For a given number of events to be required for an interim or final analysis, trial duration T+ τ can be derived using the same equation by tilting values of T and/or τ.

Section 3.4: Prediction for the To-be-randomized Subjects

As depicted in Figure 3.1b, to predict number of events during the trial, there is a cohort of subjects who were not yet randomized at t_1 and will be randomized at a known time in $(t_1, t_2]$ who can contribute to events in $(t_1, t_2]$ referred to as "e_{new}" , representing events from newly randomized subjects. Since the randomization time for a control subject is known as r_{iC} with $t_1 < r_{iC} \leq t_2$, probability of resulting in an event in interval $(t_1, t_2]$ can be calculated directly and the outer integration with respect to distribution of accrual process as shown in Appendix 3.1 is no longer needed. This approach is very different from prediction prior to trial start where randomization is a stochastic process and is modeled as uniformly distributed. Time to be randomized is now determined at r_{iC} for control subject in this cohort and time from randomization to t_2 is $t_2 - r_{iC}$. For each To-Be-Randomized subject, probability of resulting in an event can be directly calculated. Thereafter summing over each subject in this cohort from both control and treatment groups will lead to the expected number of events in $(t_1, t_2]$. After To-Be-Randomized subjects will be considered to be at risk once they are successfully randomized into the comparative double-blind phase after all protocol scheduled visits prior to it.

Appendix 3.2 describes the prediction method for this cohort of subjects during the trial. Since without censoring is a special case of with censoring, prediction with exponential censoring is derived first in Section 3.4.1 and then goes to prediction without censoring together with different parametric type of death times.

Section 3.4.1: Prediction in the Presence of Censoring

Let u_i be the time interval from randomization to end-of-study (i.e. $u_i = t_2 - r_{iC}$), for each subject in the control group. Thus, the probability of having an event for control subject i is $P[Y_c < W_c, Y_c < u_i]$ in the presence of censoring. Event of $(Y_c < W_c, Y_c < u_i)$ indicates event process occurred before the censoring process in $(t_1, t_2]$ and resulted in an event prior to t_2.

From Appendix 3.2, conditional on censoring variable, indicator variable $I(Y_c < u_i)$ can be pulled out from expectation because of independence between death time and time to censoring, which is a reasonable assumption in survival trials. Thus probability of having an event for control subject i is

$P[Y_c < W_c, Y_c < u_i] = \int_0^{u_i} f_{Y_C}(t) S_{W_C}(t) dt$, where Y_c and W_c are death time variable and censoring variable respectively, u_i is the time from randomization to t_2, $f_{Y_C}(t)$ is the density of death times and $S_{W_C}(t)$ (i.e. $\exp(-\phi_C t)$ for exponential censoring) is the survivor function for time to censoring random variable. After plugging in death time density and exponential survivor function, integrate this product with respect to time t resulting in the required probability. In case of exponential death time, $f_{Y_C}(t) = \lambda_C \exp(-\lambda_C t)$ and

$P[Y_c < W_c, Y_c < u_i] = \int_0^{u_i} \lambda_C \exp(-\lambda_C t) \exp(-\phi_C t)\, dt$

As noted above, summing over all subjects in this cohort leads to the contribution on number of events from them in time $(t_1, t_2]$. That is: $e_{new} = E(e_c) + E(e_E)$ $= \sum_{i=1}^{n_c} P[Y_c < W_c, Y_c < u_i] + \sum_{i=1}^{n_E} P[Y_E < W_E, Y_E < u_i]$

Section 3.4.2: Prediction without Censoring

With no censoring existing in the trial, $S_{W_C}(t)$ is ignored in calculating predicted probability. Hence, $\mathrm{P}[\mathrm{Y_c} < \mathrm{W_c}, \mathrm{Y_c} < \mathrm{u_i}]$ degenerates to $\mathrm{P}[\mathrm{Y_c} < \mathrm{u_i}]$, which is basically the cumulative density function for death times. See Appendix 2 for corresponding cumulative density function (CDF) for different parametric death time distributions. $e_{new} = E(e_c) + E(e_E) = \sum_{i=1}^{n_c} P[Y_c < u_i] + \sum_{i=1}^{n_E} P[Y_E < u_i]$

Section 3.4.3 When Death Time is Weibull or Another Type

Similarly, for death time other than exponential, right density of f_{Y_C} is used with exponential censoring survival function $S_{W_C}(t)$ and then integration with respect to time from 0 to u_i can result in the probability of having an event in time $(t_1, t_2]$ for subject i in this cohort.

Not every To-Be-Randomized subject would withdraw early before being randomized into the double-blind phase, only a fraction of the subjects who are ongoing in the phases prior to the double-blind phase can finish required period and then continue to be randomized into the double-blind phase at r_{iC} (with $t_1 < r_{iC} \leq t_2$) so that they can contribute to the event count in $(t_1, t_2]$. Because we only have loss to follow-up and administrative censorship in controlled clinical trial, there is no basis to assume non-constant hazard rate for time to censoring and thus only exponential censoring time is used in predicting methods in this paper throughout. However, in case having other censoring process present in the trial, other parametric censoring other than exponential can be incorporated as well. Similarly, hazard rate of death time could change over time in the trial. For example, cholesterol lowering therapies may take a year before physiologic changes are sufficient to reduce the hazard (Lipid Research Clinical Program, 1979). In this regards, parametric death times other than exponential could also be used in prediction algorithm.

Section 3.5: Prediction for the At-Risk Subjects

As illustrated in Figure 3.1b, predicting during the trial not only need to consider To-Be-Randomized subjects, but also need to determine the probability of having an event in $(t_1, t_2]$ for subjects who remained event-free right in the double-blind phase at time t_1. These subjects are considered to be in the risk set at t_1 because they potentially can have an event at any time after t_1. For these At-Risk subjects, Sections 3.5.1 and 3.5.2 illustrate the prediction algorithm in the presence of censoring and without censoring respectively. Section 3.5.3 explores prediction with Weibull death times as an example. Appendix 3.3 includes all the elements for prediction with or without censoring, and considers different parametric death times such as exponential, Weibull, Log-logistic and Log-normal.

Section 3.5.1: Prediction in the Presence of Censoring

The same considerations made in the prediction described in Sections 3.3 and 3.4 are noted here. Let random variable of time to censoring for subject i in the control group be exponentially distributed with hazard rate $\emptyset_C$.

To calculate probability of having an event in the presence of censoring by t_2 given subject is in the risk set at time t_1, two machineries are needed (Appendix 3.3). First machinery is the conditional density of having an event prior to or on t_2 conditioning on subject being in the risk set at t_1. That is, to take derivative of $P(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC})$ with respect to variable of $t_2 - r_{iC}$ when t_2 is varying from t_1 to positive infinity. The second machinery is the truncated survival function of time to censoring given time to censoring is greater than $t_1 - r_{iC}$. Excerpted from Appendix 3.3, the probability of having an event for At-Risk subject i in the control group in the presence of censoring is:

$$\begin{aligned}
&P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC}|X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC}) \\
&= E_{X_{iC}}\left[\, I(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC}) P(x_{iC} \leq W_{iC}|W_{iC} > t_1 - r_{iC}) \,\right] \\
&= \int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC})}{d(t_2 - r_{iC})} \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]} dx_{iC}
\end{aligned}$$

The probability of having an event before t_2 in the presence of censoring for At-Risk subjects is the event of $X_{iC} \leq t_2 - r_{iC}$ and $X_{iC} < W_{iC}$ given both $X_{iC} > t_1 - r_{iC}$ and $W_{iC} > t_1 - r_{iC}$, where X_{iC} and W_{iC} are exponential random variable for time to event and time to censoring for control subject i respectively. Note that although time to death are i.i.d exponential with hazard rate λ_C and time to censoring are i.i.d. exponential with hazard rate $\emptyset_C$, we put a subscript i to represent each subject in the formulation because conditional density and probabilities differ from each other due to the difference in $t_1 - r_{iC}$ resulting from different randomization time from subject to subject. Probability of event of $X_{iC} \leq t_2 - r_{iC}$ and $X_{iC} < W_{iC}$ given both $X_{iC} > t_1 - r_{iC}$ and $W_{iC} > t_1 - r_{iC}$, as in Appendix 3.3, can be expressed as the expected value of an indicator function. Conditioning on random variable of time to censoring W_{iC}, event of $X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}$ can be separated out. Then two machineries mentioned above can be multiplied together as the integrand to be integrated in the range of $(t_1 - r_{iC}, t_2 - r_{iC}]$ to get the required probability for each subject in the risk set.

For subject i in the risk set at t_1, the conditional probability accompanying with censoring is $\frac{\lambda_C[\exp[-(\lambda_C+\emptyset_C)(t_1-r_{iC})]-\exp[-(\lambda_C+\emptyset_C)(t_2-r_{iC})]]}{(\lambda_C+\emptyset_C)\exp[-(\lambda_C+\emptyset_C)(t_1-r_{iC})]}$. When $\emptyset_C$=0, the case with no censoring, this probability degenerates to $1 - \frac{\exp[-\lambda_C(t_2-r_{iC})]}{\exp[-\lambda_C(t_1-r_{iC})]}$.

Section 3.5.2: Prediction for Subjects in the Risk Set in Case There is No Censoring

Unlike the prediction carried out prior to trial start in Section 3.3, each subject in the risk set has unique randomization date, hence has varying length of time from randomization to prediction time t_1 and we do not make use of randomization accrual rate similar to what we did in predicting number of events prior to trial start. Deriving conditional probability directly for each individual and then summing all probabilities to get predicted number of events by t_2 are what we propose (Appendix 3.3). Without considering censoring, the conditional probability for subject i to have an event before t_2 given being at risk at t_1 is $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) = 1 - \frac{S_C(t_2-r_{iC})}{S_C(t_1-r_{iC})}$.

This probability degenerates to be $1 - S_C(t_2 - r_{iC})$ when $t_1 = r_{iC}$. In this case, this subject is no longer present in the risk set at t_1, but could be considered as being randomized right at t_1. The probability of having an event before t_2 should be exactly one minus the survivor probability. When plugging in exponential death time, $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})$ becomes $1 - \exp[-\lambda_C(t_2 - t_1)]$, which shows the memory-less property of exponential distribution, with which the probability is only function of $t_2 - t_1$and the time staying in the trial prior to $t_1$is fully ignored as there is no memory on it at all.

Section 3.5.3 When Death Time is Weibull or Another Type

There is no reason to assume non-constant hazard for time to censoring in clinical trial where withdrawals are non-informative with regard to death time process, but death times themselves could have non-constant hazard overtime. In case of other death time distribution, $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})$ will no longer have memory-less property for exponential death times; and $P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC})$ in the presence of censoring will be even harder to calculate. For other parametric death times, it is not easy or even possible to find the closed form for probability of having an event before t_2 for subjects in the risk set with or without censoring. However, numerical integration can easily help with calculating this probability measure. For example, consider the two-parameter Weibull distribution with hazard function $\lambda(\mathrm{t}) = \lambda\Upsilon(\lambda\mathrm{t})^{\Upsilon-1}$, $\Upsilon, \lambda > 0$. The hazard is monotone decreasing for $\Upsilon < 1$, increasing for $\Upsilon > 1$, and reduces to the constant hazard if $\Upsilon = 1$. The probability for At-Risk subject i to result in an event before t_2 is

$$E(e_{ic}) = \int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})}{d(t_2 - r_{iC})} \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]} dx_{iC}$$

$$= \int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{\lambda_C \Upsilon_C [\lambda_C x_{iC}]^{\Upsilon_C - 1} \exp[-[\lambda_C x_{iC}]^{\Upsilon_C}}{\exp[-[\lambda_C(t_1 - r_{iC})]^{\Upsilon_C}} \frac{\emptyset_C \exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]} dx_{iC} .$$

When censoring process is ignored, it degenerates to

$$E(e_{ic}) = \int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})}{d(t_2 - r_{iC})} dx_{iC} \quad =$$

$$\int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{\lambda_C \Upsilon_C [\lambda_C x_{iC}]^{\Upsilon_C - 1} \exp[-[\lambda_C x_{iC}]^{\Upsilon_C}}{\exp[-[\lambda_C(t_1 - r_{iC})]^{\Upsilon_C}} dx_{iC} .$$

Prediction for At-Risk subjects with death times in Log-logistic or Lognormal distribution with or without censoring is also explored in Appendix 3.3. Different from prediction for To-Be-Randomized subjects, all At-Risk subjects at t_1 should be evaluated to contribute to the effective number of events accumulated in $(t_1, t_2]$. The number of events from this cohort is referred as "e_{atrisk}", which is

$$E(e_c) + E(e_E) = \sum_{i=1}^{n_c} P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC}) +$$
$$\sum_{i=1}^{n_E} P(X_{iE} \leq t_2 - r_{iE}, X_{iC} < W_{iE} | X_{iE} > t_1 - r_{iE}, W_{iE} > t_1 - r_{iE})$$

or

$$E(e_c) + E(e_E) = \sum_{i=1}^{n_c} P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) +$$
$$\sum_{i=1}^{n_E} P(X_{iE} \leq t_2 - r_{iE} | X_{iE} > t_1 - r_{iE},)$$ for cases with censoring or without censoring respectively.

Section 3.6: Clinical Trial Example

During the conduct of a recently completed clinical trial (Berwaerts et al, 2015), the proposed methodology was implemented to yield accurate prediction of events. Briefly, this study evaluated the efficacy of an investigational compound compared to placebo in delay of the time to first occurrence of relapse. The study consists of 4 phases: a screening Phase (up to 3 weeks); a 17-week flexible dose open-label transition phase; a 12-week fixed dose open-label maintenance phase; and a randomized, double-blind, fixed dose, placebo-controlled relapse prevention phase of variable duration. Subjects remained in the study for as long as they were clinically stable or until the Sponsor stopped the trial.

As part of study design, it was assumed that the 12-month relapse rates for treatment and placebo will be 20% and 40%, respectively, resulting in a hazard ratio of 0.44. Approximately 196 subjects were expected to be randomized in the double-blind phase in a 1:1 ratio to either treatment or placebo in order to obtain 70 relapse events to show that treatment is significantly different from placebo at the 2-sided significance level of 0.05, with 90% power to detect a hazard ratio of 0.44. A 2-stage group-sequential design with one interim analysis was proposed to allow for early stopping if there was significant evidence of efficacy based upon the interim analysis after 60% of the projected relapse events (i.e., 42 relapse events) have occurred. It was assumed that at least 50% of subjects who enter the transition phase would discontinue the study or not meet the criteria for randomization in the double-blind Phase. To meet the expected number of 196 subjects (98 per treatment group) to be randomized in the double-blind phase, a total of 392 subjects were to be enrolled. The total number of subjects enrolled depended on the time that it would take to obtain 70 relapse events. The actual total number of subjects enrolled was 506.

Several predictions were carried out during the course of trial to help with trial monitoring. One such prediction based on data from November 29, 2013 is used for the illustration below (Figure 3.2). The study begun on April 26, 2012, first subject was randomized on November 26, 2012 and first event has occurred at December 10, 2012. Figure 3.2 illustrates the states of affairs on November 29, 2013.

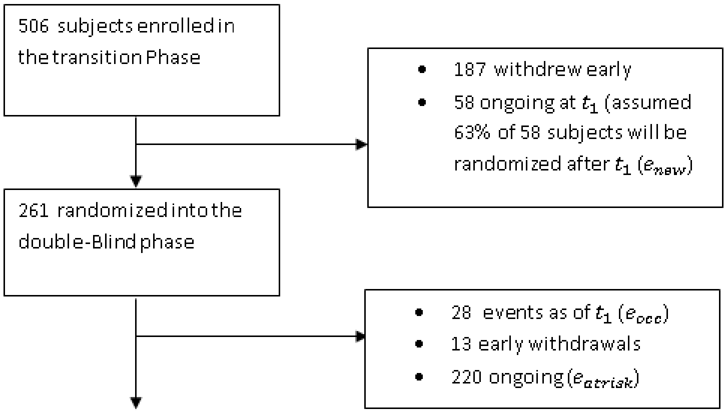

506 subjects enrolled in the transition Phase
187 withdrew early
58 ongoing at t_1 (assumed 63% of 58 subjects will be randomized after t_1 (e_{new})
261 randomized into the double-Blind phase
28 events as of t_1 (e_{occ})
13 early withdrawals
220 ongoing (e_{atrisk})

Figure 8(Fig. 3.2): Study Completion and Withdrawal

Figure 3.2: Study Completion and Withdrawal Information at Predicting Time $\mathbf{t_1}$ of November 29, 2013.

By November 29, 2013, enrollment of subjects into the transition phase has been completed, the subjects who were still in the transition/maintenance phases were the only eligible cohort to be randomized after November 29, 2013 to have event and subjects who were ongoing on November 29, 2013 could have event later on. Since 187 (37%) of the 506 enrolled subjects had withdrawn early from the transition/maintenance phase, we assume that 63% of other remaining subjects (n=58) in the transition/maintenance subjects would be randomized after t_1. Thus, a uniform random variable is generated for each of the 58 subjects and a subject will be randomized after t_1 if the uniform random variable is greater than or equal to 0.37. Therefore, among 58 subjects who were on-going in the combined transition/maintenance phases at prediction time, only 31 of them will be randomized later.

As shown in Figure 3.2, we then predict the time to achieve required number of events t_2 (with $t_2 > t_1$) based on data as of November 29, 2013 (t_1):

- 28 events occurred in the double-blind before November 29, 2013 (i.e. e_{occ}=28);
- Subjects (N=220) who were event-free in the double-blind phase on November 29, 2013. The predicted number of events before t_2 in this group is denoted as e_{atrisk};
- 63% of the subjects (n=31) who were ongoing during the transition/maintenance phases at November 29, 2013 and will be randomized after t_1. The predicted number of events in this cohort is denoted as e_{new}.

Section 3.6.1: Plotted Survival Curves at Time t_1

Before implementing prediction algorithm on cutoff date of November 29, 2014, we derive the parametric death time distribution for prediction using exponential, Weibull, Log-logistic and Lognormal distributions. Parameters were extracted after fitting data with a parametric death time distribution of interest and were then used to create parametric survivor curve over time to compare with non-parametric Kaplan-Meier (i.e. KM) curve. The parametric distribution closest to non-parametric KM plot would be considered appropriate. In order to maintain treatment information blinded, one combined group is used to extract parameters for death times instead of having treatment specific parameters. Figure 3.3 shows the KM plot along with fitted parametric death curves of exponential, Lognormal, Weibull and Log-logistic separately.

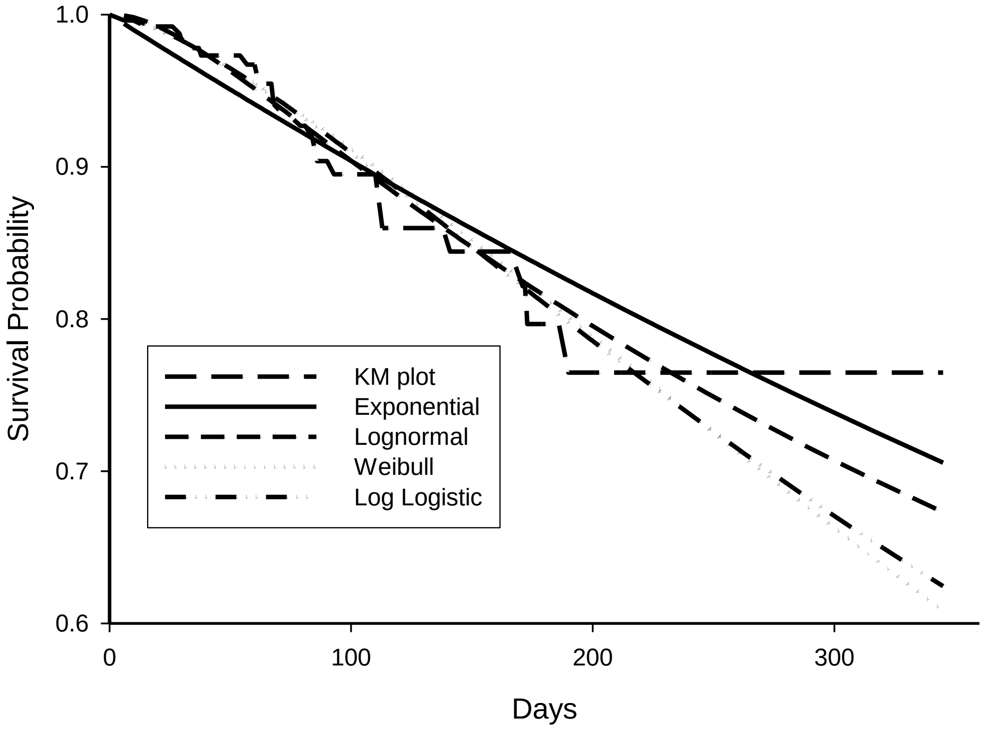


Figure 9(Fig. 3.3): KM plot and estimated parametric survivor curves at November 29

Figure 3.3: KM plot and estimated parametric survivor curves at time t_1 of November 29, 2013 for the combined group in the DB phase

In addition to the plot, we calculated the distance between a particular parametric curve and KM plot at each death time point. Suppose there are J distinct death time points in the combined group in above KM plot (multiple events can occur at the same time point), $\mathrm{s_{i,KM}}$ is the survivor probability for KM plot at ith time point while $\mathrm{s_{i,p}}$ is for a particular parametric survival curve. The sum of squared differences over all J distinct time points is summarized for death times of Exponential, Weibull, Log-logistic and Lognormal against KM plot respectively in Table 3.1.

Table 10(Tab. 3.1): Sum of squared difference between survivor curve of a parametric distribution and the KM plot

Table 3.1: Sum of squared difference between survivor curve of a parametric distribution and the KM plot

	Exponential	Weibull	Log-logistic	Lognormal
Sum of squared differences $= \sum_{i=1}^{J} (s_{i,p} - s_{i,KM})^2$	0.066	0.119	0.099	0.050

From Figure 3.4 and Table 3.1, it is difficult to choose the best parametric death time distribution to use for prediction, so all are used for prediction. This allows the prediction to yield a range of dates that could be used for trial monitoring and operational planning.

In the Section 3.6.2 details on using data from November 29, 2013 to predict the calendar time, by which 42 relapses (including 28 relapses that had occurred prior to or on November 29, 2013) can be accumulated in the double-blind phase. Parameters for each parametric death times based on the combined data were already extracted in order to do plots in Figure 3.1. The hazard parameter for exponential censoring in the double-blind phase can be obtained using the same data but by considering time to withdrawals prior to the 1st relapse and prior to November 29, 2013 as events while with the rest being censored at their relapse dates or at the cutoff date November 29, 2013. There is a reason why only early withdrawals are used as censoring events. As the main goal in this paper is to calculate probability of subject having an event prior to or on a future time and censoring process that could possibly impact this prediction is concerned. But most probably only the non-administrative censoring (i.e., early withdrawals) would have such impacts while administrative censoring won't have.

Section 3.6.2: Prediction Calendar Time to Achieve 42 Events for Interim Analysis

The prediction is carried out as follows:

- Estimate parameters for death time: on November 29, 2013 there were 28 relapses that had occurred. For subjects who were randomized but with no record of relapse are censored at either date of withdrawal or at the cutoff date. This data is used to fit exponential, Weibull, Lognormal and Log-logistic distributions, and parameters for the corresponding death time distributions can be extracted for prediction.
- Estimate exponential hazard rate for censoring: In order to estimate hazard rate for exponential censoring process, the 13 early withdrawal subjects (Figure 3.2) in the DB phase are considered as the events and others as censored. This data is fitted using an exponential distribution to get hazard rate for exponential censoring parameter $\emptyset$ in the combined group.
- Preparations for obtaining e_{new} in $(t_1, t_2]$: Using the subset data set (N=31) from those yet to-be-randomized subjects in the transition / maintenance

phases at time of November 29, 2013, we derive their randomization dates. For example, if one subject was at Week 28 visit at November 29, 2013 who will be eligible for randomization, this subject will be randomized a week later (i.e., December 6, 2013).

- Preparations for obtaining e_{atrisk} in$(t_1, t_2]$: For the 220 subjects who are already in the double-blind phase on November 29, 2013, we save their randomization dates which have occurred prior to the cutoff date for prediction.

There are 8 scenarios of predictions: Table 3.2 includes prediction results with death times of exponential, Weibull, Log-logistic, Lognormal respectively when censoring is not present; and Table 3 includes prediction results from the same set of parametric death time distributions but in the presence of censoring.

For each scenario, in order to predict t_2, the earliest time to accumulate 42 events in the double-blind phase, a date after t_1 is chosen for initial prediction. For example, we choose January 01, 2014. e_{new} and e_{atrisk} at t_2= January 01, 2014 are then calculated using algorithms in Sections 4-5 and Appendixes 2-3. If the total number of events (i.e. $e = e_{occ} + e_{new} + e_{atrisk}$) is less than 42, we then increase the date and redo calculation until the earliest date to accumulate 42 events for interim analysis is achieved.

Table 11(Tab. 3.2): Prediction of the earliest date to obtain 42 events assuming no censoring

Table 3.2: Prediction of the earliest date to obtain 42 events assuming no censoring

	Exponential		Weibull		Log-logistic		Log-normal	
e_{occ}		28		28		28		28
e_{atrisk}	t_2 =Jan20,2014	13.2	Jan10,2014	13.71	Jan11,2014	13.86	Jan12,2014	13.60
e_{new}		0.85		0.44		0.439		0.46
e by		42.05		42.14		42.29		42.06

t_2								

Table 12(Tab. 3): Prediction of the earliest date to obtain 42 events in the presence of censoring

Table 3: Prediction of the earliest date to obtain 42 events in the presence of censoring

	Exponential		Weibull		Log-logistic		Log-normal	
e_{occ}		28		28		28		28
e_{atrisk}	t_2=Feb 6,2014	12.83	Jan10,2014	13.88	Jan11,2014	13.69	Jan 13,2014	13.74
e_{new}		1.22		0.45		0.43		0.46
e by t_2		42.05		42.33		42.12		42.20

Results of the prediction ranged from Jan 10, 2014 (using Weibull death times with/without censoring) to Feb 6, 2014 (exponential death time in the presence of censoring). For each death time, predicted date of t_2 for the case with censoring is later than or the same as the date using the same death time distribution but without censoring. This is understandable, because with time to censoring competing with process of time to event, the time to get required events will be delayed. In our data, we actually only have 13 early withdrawals out of total 261 randomized subjects. So the time to censoring barely impacted the prediction dates.

In our example, prediction based on exponential model differs from predictions using other models, while exponential is easiest one among all prediction and wildly used in design and monitoring survival trials. This suggests that one cannot rely on one particular parametric model. In the actual study, the required 42 events needed for interim analysis was observed on January 24. Based on the prediction, the study team was able to plan appropriately and external Statistical Support Group (supporting the Independent Data Monitoring Committee) was ready to go as soon the requisite time point was reached. Figure 3.4 below depicts predicted total number of events from the prediction carried out on November 29 2013 in the absence or presence of censoring until the 42 events needed for interim analysis are achieved, compared with the actual curve for total number of events the trial ended up with (solid line). The upper and lower plots include depict predictions in the absence and in the presence of censoring respectively.

An earlier prediction with data cutoff of October 16, 2013 (when only 20 events had been observed) actually resulted in predicted date range from December 24, 2013 to January 20, 2014 (Figure 3.5), which was less accurate than predictions done one month later on November 29, 2013 (Figure 3.4).

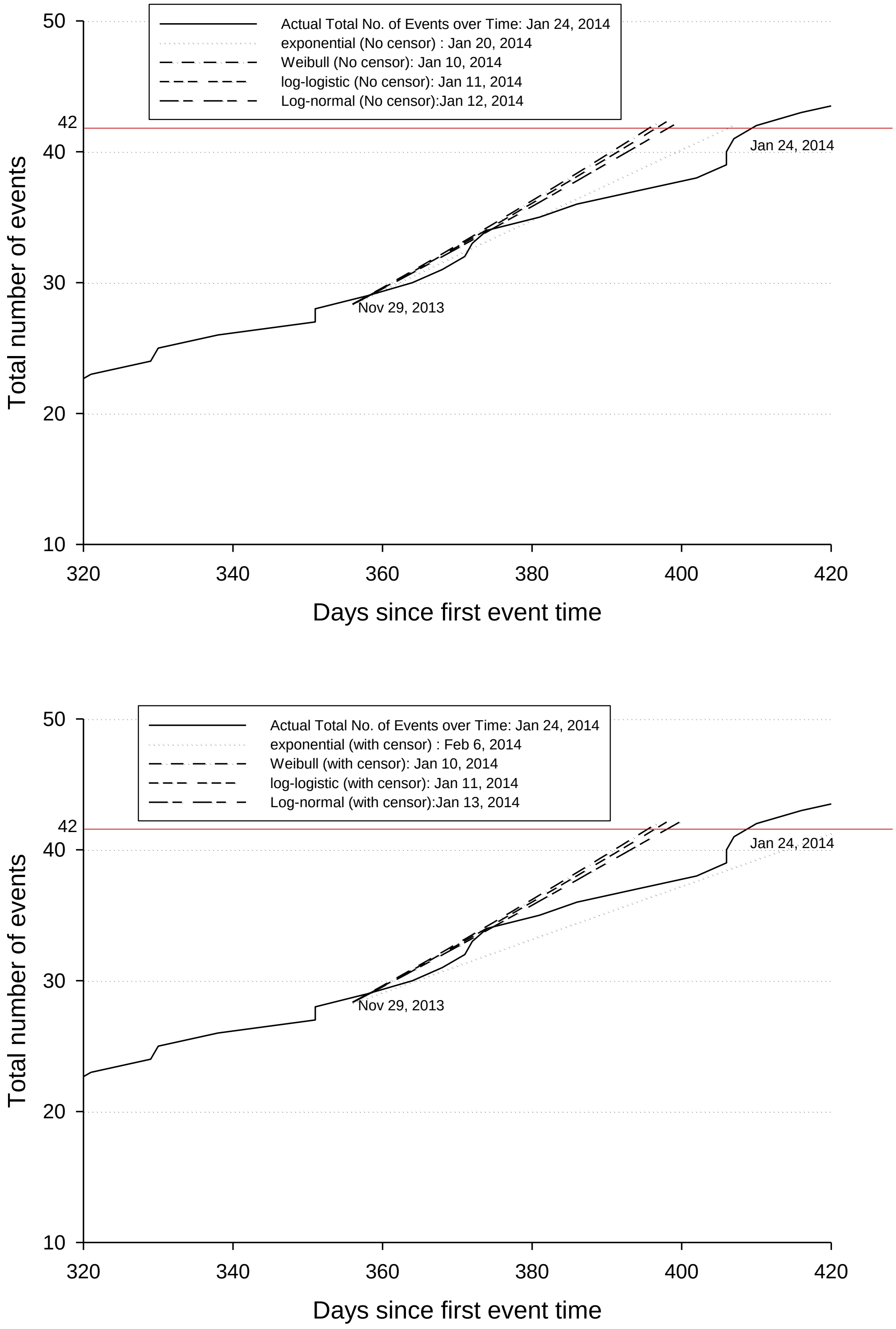
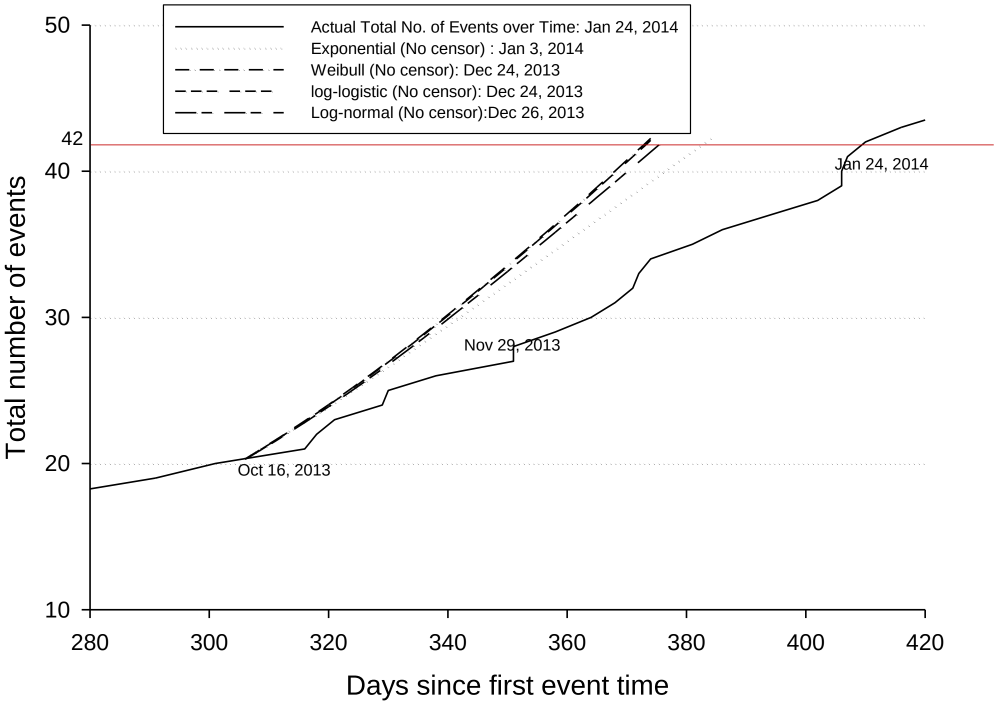


Figure 10(Fig. 3.4): Total number of events over time from prediction time November 2013

Figure 3.4: Total number of events over time from prediction time t_1=29

November 2013 until reaching 42 events. The upper and lower plots include predictions in the absence and in the presence of censoring respectively.



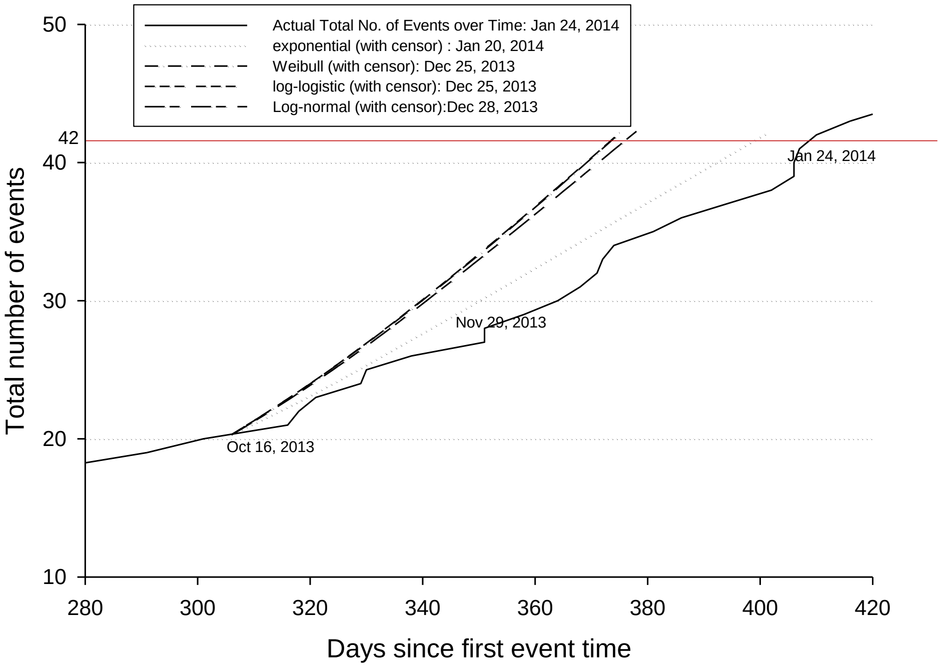


Figure 11(Fig. 3.5): Total number of events over time from prediction time October 2013

Figure 3.5: Total number of events over time from prediction time t_1=16 October 2013 until reaching 42 events. The upper and lower plots include predictions in the absence and in the presence of censoring respectively.

Section 3.7: Discussion

This paper extends Whitehead (2001) to include prediction in the presence of censoring prior to trial start. Inspired by the need to know when a certain number of events would be observed during the trial, we develop methodologies to carry out prediction during the trial with or without censoring using different parametric death time distributions. Technical details (Appendix 3.1-3.3) are inspired by statistical appendix in Rubinstein, Gail and Santner (1981). The key is that in the presence of censoring, the integrand part of this probability can be separated into two parts because of the independence between death time and time to censoring. For subjects who will be randomized at a given date in $(t_1, t_2]$, one part is the unconditional density of death time and the other is the unconditional survivor function for censoring time; for subjects who are already randomized and in the at-risk set at prediction time t_1, one part is the conditional density of death time and the other is the conditional survivor function of censoring time given both death time and censor time are greater than $t_1 - r_{iC}$. For prediction during the trial, given t_2, integration range (the time interval in which this subject will result in an event) for each individual is known and thus the probability of resulting in an event can be obtained directly. Summing up probability over all subjects in corresponding cohort will obtain the expected number of events in the interval of interest because expectation of an indicator function equals its probability and expectation of sum equals the sum of expectations. For prediction prior to trial start, an additional integration with respect to randomization accrual variable is needed to obtain the expected number of events

(Appendix 3.1) prior to calendar time $T + \tau$.

Methods derived here are both easy to understand and easy to implement. Knowing the possible calendar time for interim analysis ahead of time makes trial planning much easier, and needed resources can be deployed in a timely manner such as getting database ready to be locked for final analysis. Successful prediction during the course of an actual trial in Section 6 corroborated this claim. Before study start, the prediction had been based on exponential distribution and study start assumptions suggesting some time during the third quarter of 2014. The prediction work at later times allowed the team to adjust timelines based on actual trial data. A more accurate prediction is needed for trial management, especially for a globally-managed trial involving many patients, personnel, and functions. The resulting prediction suggested a first quarter interim analysis.

The prediction algorithm used combined treatment information so there was no need to unblind the treatment arms. Assumption about treatment group differences have to be made and this may affect the precision of the prediction. But our trial experience showed that prediction based on a combined group is good enough for trial management. Initial trial prediction is based on the same assumptions made for sample size calculation, and can be enhanced with actual accumulated data. Our methods of using a series of parametric distributions for single predicted time contrasts with other methods based on simulating empirical distribution of predicted target time t_2 based on posterior sampled parameters as illustrated in Bayesian methods. Although the latter method has also been used to obtain an interval around the prediction time, extra sampling/prediction errors will be added for an algorithm which already includes uncertainty from prior and MCMC sampling for incomplete data and posteriors. Nonparametric prediction using Kaplan-Meier estimator to extrapolate the survival probability into the future together with Bayesian bootstrapped prediction intervals has also been proposed by Ying, Heitjan and Chen (2004); but was shown to less accurate than predictions using Bayesian parametric prediction by the same group of authors (Ying and Heitjan, 2008). Detailed comparisons between these various Bayesian methods using parametric or non-parametric event times with our method can be the subject of future research.

References

Bagiella E, Heitjan DF. Predicting analysis times in randomized clinical trials. *Statistics in Medicine* 2001; 20:2055-2063.

Berwaerts J, Liu Y, Gopal S, Nuamah I, Xu H, Savitz A, Coppola D, Schotte A, Remmerie B, Maruta N, Hough D. Placebo-controlled relapse prevention study of the 3-month formulation of paliperidone palmitate for schizophrenia: a randomized clinical trial. *JAMA Psychiatry*. 2015. doi:10.1001/ jamapsychiatry.2015.0241. Published online March 29, 2015.

Collett D. *Modeling Survival Data in Medical Research*. London: Chapman and Hall. 1994.

Donovan JM, Elliott MR, Heitjan DF. Predicting event times in clinical trials when treatment arm is masked. *Journal of Biopharmaceutical Statistics* 2006; 16:343-356.

Donovan JM, Elliott MR, Heitjan DF. Predicting event times in clinical trials when randomization is masked and blocked. *Clinical trials* 2007; 4:481-490.

FDA Guidance for industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf.

George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Disease* 1974; 27:15-29.Lee ET, Desu MM, Gehan EA. A Monte Carlo study of the power of some two-sample tests. *Biometrika* 1975; 62:425-431.

Lipid Research Clinics Program. The Coronary Primary Prevention Trial: design and implementation. *Journal of Chronic Disease* 1979; 32:609-631.

Mantel N. Evaluation of survival data and two rank order statistics arising in its consideration. *Cancer Chemotherapy Rep* 1966; 50:163-170.

O’Brien PC, Fleming T R. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549-56.

Pasternak BS, Gilbert HS. Planning the duration of long-term survival time studies designed for accrual by cohorts. *Journal of Chronical Disease* 1971; **27**:681-700.

Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease* 1981; 34: 469-479.

Wang J, Ke C, Jiang Q, Zhang C, Snapinn, S. Predicting analysis time in event-driven clinical trials with event-reporting lag. *Statistics in Medicine* 2012; 31: 801-811.

Whitehead, J. ‘Predicting the duration of sequential survival studies’, *Drug Information Journal* 2001; 35: 1387-1400.

Ying G, Heitjian DF, Chen T. Nonparametric prediction of event times in randomized clinical trials. *Clinical trials* 2004; 1:352-361.

Ying G, Heitjian DF. Weibull prediction of event times in clinical trials. *Pharmaceutical Statistics* 2008, **7**:107-120.

Zhang X, Long Q. Stochastic modeling and prediction for accrual in clinical trials. *Statistics in Medicine* 2010; 29:649-658

Zhang X, Long Q. Joint monitoring and prediction of accrual and event times in clinical trials. *Biometrical Journal* 2012a; 54:735-749

Zhang X, Long Q. Modeling and prediction of subject accrual and event times in clinical trials: a systematic review. *Clinical Trials* 2012b, 9: 681-688.

Appendix 3.1: Prediction prior to trial start with exponential death time and exponential censoring

Assuming that patients are uniformly randomized into an interval [0, T] in unit of month, the total number of subjects entering the DB phase N=n_E +n_C will be aT in total with recruitment rate of a per month over the T month accrual. With randomization ratio A:1 of treatment group (n_E) to control group (n_C), then the expected recruitment in T months for treatment and control groups respectively are: $E[n_E] = \frac{A}{A+1}aT$ and $E[n_C] = \frac{1}{A+1}aT$. Given N, the patient's entry times will be independently and identically distributed (i.i.d.) uniformly over [0, T]. Therefore, with follow-up time τ, the times from randomization to end-of-study (EOS) will be i.i.d. uniform over [τ, T+ τ] (Figure 1a).

Given a time interval u from randomization onto control group end-of-study, the probability that this entry will result in an event is:

$P[Y_c < W_c, Y_c < u] = E[\, I(Y_c < W_c, Y_c < u)] = E[\; E[\, I(Y_c < W_c, Y_c < u)|W_c]\,]$

=E[$I(\, Y_c < u)$ $E[\, I(Y_c < W_c)|W_c]\,]$ because of independence between W_c and Y_c

=E[$I(\, Y_c < u)$ $S_{W_C}(u))$]

$$= \int_0^u f_{Y_C}\,(t) S_{W_C}(t)dt$$

$S_{W_C}(u)$ is the survivor function of time to censoring variable while W_c is exponentially distributed with constant hazard ϕ_C, that is $S_{W_C}(t) = \exp(-\,\phi_C t)$. f_{Y_C} is the probability density function of time to event in the control group, which has constant hazard λ_C with density function $f_{Y_C} = \lambda_C \exp(\,\lambda_C)$. Plugging the density and survivor functions in, we obtain,

$P[Y_c < W_c, Y_c < u] = \int_0^u \lambda_C \exp(-\,\lambda_C t) \exp(-\,\phi_C t)\, dt =$
$\frac{\lambda_C}{\lambda_C + \phi_C}[\, 1 - \exp[-(\,\lambda_C + \,\phi_C)u]\,]$

Similar definitions hold for the treatment group, we have

$P[Y_E < W_E, Y_E < u] = \int_0^u \lambda_E \exp(-\,\lambda_E t) \exp(-\,\phi_E t)\, dt =$
$\frac{\lambda_E}{\lambda_E + \phi_E}[\, 1 - \exp[-(\,\lambda_E + \,\phi_E)u]\,]$

During the T+ τ months of trial duration, given n_C subjects randomized into the control group, the expected number of events in this group is as follows:

$E(e_c|n_C) = n_C P(\text{event on control})= n_C E[\, E[\, I(Y_c < W_c, Y_c < u)|u\,]\,]$
=
$n_C \int_\tau^{T+\tau} P(event\ on\ control|time\ from\ randomization\ to\ EOS\ being\ u)g(u)du$

where $g(u)$ is the density of u

$= n_C \int_\tau^{T+\tau} \frac{\lambda_C}{\lambda_C + \phi_C}[\, 1 - \exp[-(\,\lambda_C + \,\phi_C)u]\,]\frac{1}{T}du$

$= \frac{n_C\,\lambda_C}{T\,(\,\lambda_C + \phi_C)}\,[\,T + \frac{\exp[-(\,\lambda_C + \phi_C)(T+\tau)] - \exp[-(\,\lambda_C + \phi_C)\tau]}{\lambda_C + \phi_C}]$

So $E(e_c) = E[\, E(e_c|n_C)\,]$

$= \frac{E(n_C)\,\lambda_C}{T\,(\,\lambda_C + \phi_C)}\left[T + \frac{\exp[-(\,\lambda_C + \phi_C)(T+\tau)] - \exp[-(\,\lambda_C + \phi_C)\tau]}{\lambda_C + \phi_C}\right]$

$$= \frac{a\,\lambda_C}{(A+1)\,(\lambda_C + \phi_C)} \left[T + \frac{\exp[-(\lambda_C + \phi_C)(T+\tau)] - \exp[-(\lambda_C + \phi_C)\tau]}{\lambda_C + \phi_C} \right]$$

And

$$E(e_E) = E[\,E(e_E|n_E)\,] = \frac{aA\,\lambda_E}{(A+1)\,(\lambda_E + \phi_E)} \left[T + \frac{\exp[-(\lambda_E + \phi_E)(T+\tau)] - \exp[-(\lambda_E + \phi_E)\tau]}{\lambda_E + \phi_E} \right]$$

Thus,

$$E(e) = E(e_C) + E(e_E) = \frac{aT\,\lambda_C}{(A+1)\,(\lambda_C + \phi_C)} + \frac{aA\,T\lambda_E}{(A+1)\,(\lambda_E + \phi_E)} + \frac{a\,\lambda_C\,[\exp[-(\lambda_C + \phi_C)(T+\tau)] - \exp[-(\lambda_C + \phi_C)\tau]]}{(A+1)\,(\lambda_C + \phi_C)^2} + \frac{aA\,\lambda_E[\exp[-(\lambda_E + \phi_E)(T+\tau)] - \exp[-(\lambda_E + \phi_E)\tau]]}{(A+1)\,(\lambda_E + \phi_E)^2},$$

whenever there is no censoring, $\phi_C = \phi_E = 0$, the expected number of new randomized subjects degenerates to:

$$E(e) = \frac{aT}{(A+1)} + \frac{aAT}{(A+1)} + \frac{a\,[\exp[-\lambda_C(T+\tau)] - \exp[-\lambda_C\,\tau]]}{(A+1)\,\lambda_C} + \frac{aA[\exp[-\lambda_E(T+\tau)] - \exp[-\lambda_E\,\tau]]}{(A+1)\,\lambda_E}$$

Appendix 3.2: Prediction for To-Be-Randomized subjects who will be randomized at a known time between t_1 and t_2

At time t_1, we are interested in calculating the probability of resulting in an event prior to or on t_2 for those subjects who will be randomized between t_1 and t_2 $(t_1 < r_{iC} \le t_{2)}$. Since the randomization time for a control subject is known as r_{iC} with $t_1 < r_{iC} \le t_2$, probability of resulting in an event in interval $(t_1, t_2]$ can be calculated directly and the outer integration as in Appendix 3.1 with respect to distribution of accrual process is no longer needed.

u_i is the time interval from randomization onto control group end-of-study (i.e., $u_i = t_2 - r_{iC}$), and the probability that this subject will result in an event is:

$P[Y_c < W_c, Y_c < u_i] = E[I(Y_c < W_c, Y_c < u_i)] = E[\ E[I(Y_c < W_c, Y_c < u_i)|W_c]\ ]$

=E[$I(Y_c < u_i)$ $E[I(Y_c < W_c)|W_c]]$ because of independence between W_c and Y_c

=E[$I(Y_c < u_i)$ $S_{W_C}(u_i))$]

$= \int_0^{u_i} f_{Y_C}(t) S_{W_C}(t)dt = \int_0^{t_2 - r_{iC}} f_{Y_C}(t) S_{W_C}(t)dt$

Exponential censoring is used in prediction with survivor function $S_{W_C}(t) = \exp(-\phi_C t)$. For exponential death times, density of $f_{Y_C}(t)$ is already given above in Appendix 3.1. The following are the death time densities when death times are distributed with Weibull, log-normal or log-logistic function respectively.

Weibull: $f_{Y_C}(t) = \gamma_c \alpha_c t^{\gamma_c - 1} \exp(-\alpha_c t^{\gamma_c})$ where $\sigma_c = 1/\gamma_c$ and $\alpha_c = \exp(-\mu_c/\sigma_c)$

Log-logistic: $f_{Y_C}(t) = \frac{\alpha_c \gamma_c t^{\gamma_c - 1}}{(1+\alpha_c t \gamma_c)^2}$ where $\gamma_c = 1/\sigma_c$ and $\alpha_c = \exp(-\mu_c/\sigma_c)$

Log-normal: $f_{Y_C}(t) = \frac{1}{\sqrt{2\pi\sigma_c t}} \exp(-\frac{1}{2}(\frac{\log(t) - \mu_c}{\sigma_c})^2)$

Therefore, $P[Y_c < W_c, Y_c < u_i]$ $(i.e., E[I(Y_c < W_c, Y_c < u_i)]$) for death times of exponential, Weibull, Log-logistic and Log-normal are respectively:

Exponential: $\int_0^{u_i} \lambda_C \exp(-\lambda_C t) \exp(-\phi_C t)\, dt$

Weibull: $\int_0^{u_i} \gamma_c \alpha_c t^{\gamma_c - 1} \exp(-\phi_C t)\, dt$

Log-logistic: $\int_0^{u_i} \frac{\alpha_c \gamma_c t^{\gamma_c - 1}}{(1+\alpha_c t \gamma_c)^2} \exp(-\phi_C t)\, dt$

Log-normal: $\int_0^{u_i} \frac{1}{\sqrt{2\pi\sigma_c t}} \exp(-\frac{1}{2}(\frac{\log(t) - \mu_c}{\sigma_c})^2) \exp(-\phi_C t)\, dt$

In case of no censoring, $\int_0^{u_i} f_{Y_C}(t) S_{W_C}(t)dt$ degenerates to $\int_0^{u_i} f_{Y_C}(t)dt = S(u_i)$, the cumulative density function of respective death time distribution. These are:

Exponential: $\int_0^{u_i} \lambda_C \exp(-\lambda_C t)\, dt = \exp(-\lambda_C u_i)$

Weibull: $\int_0^{u_i} \gamma_c \alpha_c t^{\gamma_c - 1}\, dt = \exp(-\alpha_c {u_i}^{\gamma_c})$

Log-logistic: $\int_0^{u_i} \frac{\alpha_c \gamma_c t^{\gamma_c - 1}}{(1+\alpha_c t^{\gamma_c})^2} dt = \frac{1}{1+\alpha_c {u_i}^{\gamma_c}}$

Log-normal: $\int_0^{u_i} \frac{1}{\sqrt{2\pi\sigma_c t}} \exp(-\frac{1}{2}(\frac{\log(t) - \mu_c}{\sigma_c})^2)\, dt = 1 - \Phi(\frac{\log(u_i) - \mu_c}{\sigma_c})$

In the case of no censoring, the probability of having an event in interval $(t_1, t_2]$ after randomization equals the CDF function with time length u_i, where u_i varies and depends on when this subject will be randomized in $(t_1, t_2]$. Closed form for individual CDF is provided as above. In the case where censoring is present, this probability can be obtained by numerical integration with formulas provided.

$e_{new} = E(e_c) + E(e_E) = \sum_{i=1}^{n_c} P[Y_c < W_c, Y_c < u_i] + \sum_{i=1}^{n_E} P[Y_E < W_E, Y_E < u_i]$

Appendix 3.3: Prediction for At-Risk subjects

We first work on the conditional probability of a subject having an event before t_2 in the DB phase, given that the subject was still in the risk set at time t_1.

$P(X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC})$
$= E[\, I((X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC}))\,]$
$= E[\, E[\, I\big((X_{iC} \leq t_2 - r_{iC}, X_{iC} < W_{iC} | X_{iC} > t_1 - r_{iC}, W_{iC} > t_1 - r_{iC})\big|\, W_{iC}\,]\,]$
$= E_{X_{iC}}[\, E_{W_{iC}}[\, I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) I(X_{iC} < W_{iC} | W_{iC} > t_1 - r_{iC})]\,]$
$= E_{X_{iC}}[\, I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})\, E_{W_{iC}}[\, I(X_{iC} < W_{iC} | W_{iC} > t_1 - r_{iC})\,]\,]$
$= \; E_{X_{iC}}[\, I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) P(x_{iC} \leq W_{iC} | W_{iC} > t_1 - r_{iC})\ ]$

Note that $P(x_{iC} \leq W_{iC} | W_{iC} > t_1 - r_{iC}) = S_{W_c}(x_{iC} | W_{iC} > t_1 - r_{iC})$ is indeed the conditional survivor function for censoring random variable, which can be calculated by integrating of conditional density function over constrained interval $[t_1 - r_{iC}, +\infty)$. The conditional density function is: $f(w_{iC} | W_{iC} > t_1 - r_{iC}) = \frac{g(w_{iC})}{S_{W_C}(t_1 - r_{iC})}$, where $g(w_{iC})$ is the same as unconditional density function for random variable w_{iC} , that is $f(w_{iC}) = \emptyset_C \exp(-\emptyset_C w_{iC})$, but restricted on the set of $(t_1 - r_{iC}, \infty)$. For exponential censoring, this becomes:

$P(W_{iC} | W_{iC} > t_1 - r_{iC}) \; = \int_{x_{iC}}^{\infty} \frac{g(w_{iC})}{S_{W_C}(t_1 - r_{iC})} \; = \int_{x_{iC}}^{\infty} \frac{\emptyset_C \exp(-\emptyset_C w_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]} dw_{iC} \; = \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]}$ with $x_{iC} \in (t_1 - r_{iC}, +\infty)$.

Plugging in the conditional survivor function for exponential censoring,

$E_{X_{iC}}[\, I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) P(x_{iC} \leq W_{iC} | W_{iC} > t_1 - r_{iC})\ ]$
$= \; E_{X_{iC}}[\, I(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]}\,]$
$= \int_{t_1 - r_{iC}}^{t_2 - r_{iC}} \frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})}{d(t_2 - r_{iC})} \frac{\exp(-\emptyset_C x_{iC})}{\exp[-\emptyset_C(t_1 - r_{iC})]} dx_{iC}$

In order to calculate conditional probability of having an event before t_2 for subjects who are still at risk at time t_1, we have to get the derivative of the conditional CDF of death time with respect to $t_2 - r_{iC}$ provided that subject is at risk set at t_1, i.e., $\frac{dP(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})}{d(t_2 - r_{iC})}$, which can be obtained by taking derivative of conditional probability of $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})$ with respect to time length from randomization to t_2, that is $t_2 - r_{iC}$. We calculate $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC})$ for each parametric death times first.

$P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) = \frac{P(t_1 - r_{iC} \leq X_{iC} < t_2 - r_{iC})}{P(X_{iC} > t_1 - r_{iC})} = \frac{S_C(t_1 - r_{iC}) - S_C(t_2 - r_{iC})}{S_C(t_1 - r_{iC})}$
$= 1 - \frac{S_C(t_2 - r_{iC})}{S_C(t_1 - r_{iC})}$, where $S_C(t_1 - r_{iC})\; and\; S_C(t_2 - r_{iC})$ are unconditional survivor function at time $t_1 - r_{iC}$ and $t_2 - r_{iC}$ respectively.

Exponential: $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) = 1 - \frac{\exp[-\lambda_C(t_2 - r_{iC})\,]}{\exp[-\lambda_C(t_1 - r_{iC})}$

Weibull: $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) = 1 - \frac{\exp[-\alpha_c(t_2 - r_{iC})^{\gamma_c}\,]}{\exp[-\alpha_c(t_1 - r_{iC})^{\gamma_c}\,]}$

Log-logistic: $P(X_{iC} \leq t_2 - r_{iC} | X_{iC} > t_1 - r_{iC}) =$

$$1-\frac{\frac{1}{1+\alpha_c(t_2-r_{iC})^{\gamma_c}}}{\frac{1}{1+\alpha_c(t_1-r_{iC})^{\gamma_c}}}=1-\frac{1+\alpha_c(t_1-r_{iC})^{\gamma_c}}{1+\alpha_c(t_2-r_{iC})^{\gamma_c}}$$

Lognormal: $P(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC}) = 1 - \frac{1-\Phi(\frac{\log(t_2-r_{iC})-\mu_C}{\sigma_C})}{1-\Phi(\frac{\log(t_1-r_{iC})-\mu_C}{\sigma_C})}$

Taking derivative with respect to $t_2 - r_{iC}$, provided that $t_1 - r_{iC}$ is a fixed value for subjects still at risk at time t_1.

$\frac{dP(X_{iC}\leq t_2-r_{iC}|X_{iC}>t_1-r_{iC})}{d(t_2-r_{iC})}$ for death times of exponential, Weibull, log-logistic and lognormal are then respectively calculated as the follows:

Exponential: $\frac{dP(X_{iC}\leq t_2-r_{iC}|X_{iC}>t_1-r_{iC})}{d(t_2-r_{iC})} = \frac{\lambda_C\exp[-\lambda_C(t_2-r_{iC})\,]}{\exp[-\lambda_C(t_1-r_{iC})}$

Weibull: $\frac{dP(X_{iC}\leq t_2-r_{iC}|X_{iC}>t_1-r_{iC})}{d(t_2-r_{iC})} = \frac{\alpha_c(t_2-r_{iC})^{\gamma_c-1}\exp[-\alpha_c(t_2-r_{iC})^{\gamma_c}\,]}{\exp[-\alpha_c(t_1-r_{iC})^{\gamma_c}\,]}$

Log-logistic: $\frac{dP(X_{iC}\leq t_2-r_{iC}|X_{iC}>t_1-r_{iC})}{d(t_2-r_{iC})} = \frac{[1+\alpha_c(t_1-r_{iC})^{\gamma_c}]*\gamma_c\alpha_c(t_2-r_{iC})^{\gamma_c-1}}{[1+\alpha_c(t_2-r_{iC})^{\gamma_c}]^2}$

Lognormal: $\frac{dP(X_{iC}\leq t_2-r_{iC}|X_{iC}>t_1-r_{iC})}{d(t_2-r_{iC})} = \frac{\exp(-\left(\frac{\log(t_2-r_{iC})-\mu_C}{\sigma_C}\right)^2)/(2*\sqrt{2\pi})}{1-\Phi\left(\frac{\log(t_1-r_{iC})-\mu_C}{\sigma_C}\right)}(\frac{1}{\sigma_C}*\frac{1}{t_2-r_{iC}})$

Combining conditional death time density and conditional survivor function for censoring, we have the following form of conditional probability for subjects in the risk set at t_1 to result in an event in interval $(t_1, t_2]$:

For exponential:

$E_{X_{iC}}\,[\,I(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC})P(x_{iC} \leq W_{iC}|W_{iC} > t_1 - r_{iC})\ ]$

$= \int_{t_1-r_{iC}}^{t_2-r_{iC}} \lambda_C\exp(\,-\lambda_C[x_{iC}\ -(t_1 - r_{iC})]\,)\frac{\exp(-\emptyset_C x_{iC}\,)}{\exp[-\emptyset_C(t_1-r_{iC})]}dx_{iC}$

When $t_1 - r_{iC} = 0$, i.e., a subject is randomized at time t_1, this integration degenerates to the unconditional case as in Appendix 1. That is:

$\int_0^{t_2-r_{iC}} \lambda_C\exp(-\lambda_C x_{iC}\,)\exp(-\emptyset_C x_{iC}\,)\,dx_{iC}$, which is consistent with what we derived in Appendix 1. This further confirms the correctness of our derivation.

For Weibull, log-logistic and lognormal death times, there is no closed form for this complicated integration.

Weibull:

$E_{X_{iC}}\,[\,I(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC})P(x_{iC} \leq W_{iC}|W_{iC} > t_1 - r_{iC})\ ]$

$= \int_{t_1-r_{iC}}^{t_2-r_{iC}} \frac{\alpha_c x_{iC}{}^{\gamma_c-1}\exp[-\alpha_c x_{iC}{}^{\gamma_c}\,]}{\exp[-\alpha_c(t_1-r_{iC})^{\gamma_c}\,]}\frac{\exp(-\emptyset_C x_{iC}\,)}{\exp[-\emptyset_C(t_1-r_{iC})]}dx_{iC}$

Log-logistic: $\int_{t_1-r_{iC}}^{t_2-r_{iC}} \frac{[1+\alpha_c(t_1-r_{iC})^{\gamma_c}]*\gamma_c\alpha_c x_{iC}{}^{\gamma_c-1}}{[1+\alpha_c x_{iC}{}^{\gamma_c}]^2}\frac{\exp(-\emptyset_C x_{iC}\,)}{\exp[-\emptyset_C(t_1-r_{iC})]}dx_{iC}$

Lognormal: $\int_{t_1-r_{iC}}^{t_2-r_{iC}} \frac{\exp(-\left(\frac{\log(x_{iC}\,)-\mu_C}{\sigma_C}\right)^2)/(2*\sqrt{2\pi})}{1-\Phi\left(\frac{\log(t_1-r_{iC})-\mu_C}{\sigma_C}\right)}(\frac{1}{\sigma_C}*\frac{1}{t_2-r_{iC}})\frac{\exp(-\emptyset_C x_{iC}\,)}{\exp[-\emptyset_C(t_1-r_{iC})]}dx_{iC}$

In the case of no censoring, there is no need to go through the above process of taking derivative, times conditional survivor function for censoring variable, and then integrating the product integrand back from $t_1 - r_{iC}$ to $t_2 - r_{iC}$, simply $P(X_{iC} \leq t_2 - r_{iC}|X_{iC} > t_1 - r_{iC})$ is already the correct probability of resulting in an event in interval $(t_1, t_2]$ for subjects at risk at t_1.

$E(e_c) = E[\, E(e_{ic}|subject\ i\ treated\ with\ control\ and\ in\ the\ risk\ set)]]$

And $e_{at\ risk} = E(e) = \sum_{i=1}^{n_c} E(e_{ic}) + \sum_{i=1}^{n_E} E(e_{iE})$

Chapter 4

Planning a Comparative Group Sequential Clinical Trial with Loss to Follow-up and a Period of Continued Observation

(being reviewed by *Statistics in Biopharmaceutical Research*)

Abstract: This paper is motivated by Rubinstein, et al., (1981) and Kim and Tsiatis (1990) to provide a way in designing group sequential trials analyzed using logrank test for comparing survival under two treatments with loss to follow-up and a period of continued observation, which are frequently encountered in Phase II/III clinical trials. A method is developed to calculate the length of accrual period to assure a desired power for given control group median time to event, hazard ratio, length of the period of continued observation, information time of analyses and times of analyses, hazard rate of time to censoring and significance level. The results show that, similar to trials with fixed duration (Rubinstein, et al. 1981), introducing a period of continued observation after the end of patient accrual period reduces the total number of patients required to detect treatment effect substantially. Assuming both time to event and time to censoring (loss to follow-up) are exponential, the estimator of log hazard ratio (placebo vs. treatment) is used to test the null hypothesis of equality in survival distributions between treatment and placebo groups. Tables are created in which total trial duration are calculated for a wide range of cases for O'Brien and Fleming (1979), Pocock (1977) and Wang and Tsiatis (1987) efficacy upper boundaries, respectively. For the same accrual rate, three different curves are depicted to show the impacts of time to censoring and a period of continued observation on accrual time to ensure power in respective group sequential settings.

Key Words: Survival Trials; A period of Continued Observation; Group Sequential Design.

Section 4.1: Introduction

In clinical trials with survival data, patients are accrued in an accrual period, during which patients are screened if the inclusion and exclusion criteria are met, may or may not be required to go through a phase or a couple of phases before randomization, then all patients who meet randomization criteria can be randomized to either treatment or control group in a ratio of $A:1$ (treatment versus placebo). The accrual period in this article starts from the first subject being randomized until the last subject is randomized, and the rate of accrual is assumed to be uniform. The accrual period is followed by a period of "continued observation", in which all subjects in the trial are still exposed to study medication (i.e., treatment or placebo). After being randomized into the randomization phase until the end of the study (i.e., including both the accrual period and the continuation period), subjects can have failure, or loss to follow-up (due to loss to contact, subject consent, due to adverse event or other reasons), or remain event-free at the time of study termination. Except for subjects who have failed, all other subjects are considered to be censored in the randomization phase. The logrank statistic, also viewed as a time stratified Cochran-Mantel-Haenszel test, is the hypothesis test to compare the survival distribution of two groups, which is non-parametric and appropriate to use when the data are right-censored and the censoring is independent of the failure process. The test was proposed by Nathan Mantel (1966) and was named by Richard and Julian Peto (1972). Logrank test statistic is constructed by computing the difference between the observed and expected numbers of events in one of the two groups at each unique observed event time and then summing this difference over event time points so that a measure for the overall summary across event time points is obtained to evaluate two survival distributions in their entirety. The logrank statistic can also be derived as the score test for the Cox Proportional Hazard model (Cox, David R, 1972) comparing two groups. Based on efficiency of the score test, it is therefore asymptotically equivalent to the likelihood ratio test statistic if the proportional hazard model holds, whereas exponential failure time is a special case of the proportional hazard model. George and Desu (1974) proved that the total duration is minimized when we continue to randomize subjects into the randomization phase until the end of the trial (i.e., no period of continued observation after accrual period). Rubinstein, Gail and Santer (1981) explored the impact of a period of continued observation on the number of patients to be accrued to ensure a required statistical power and found that although total duration of the trial is increased a little as compared with that of the case with no continued observation period, accrual time could be reduced substantially as high as 50% or more after introducing a period of continued observation. Besides substantial cost saving because of reducing the required number of patients to be randomized, regulatory agencies normally challenge survival trials without a reasonable period of continued observation especially when a large cohort of patients get randomized right close to study termination. This is because this cohort of patients had not been exposed to the study medication long enough to differentiate the treatment-placebo difference before trial termination and, hence, how this cohort contributes to overall drug effect is questionable. Of note, both George and Desu (1974) and Rubinstein, Gail and Santer (1981) only focused on fixed sample design and similar investigations under group sequential setting are not yet done.

As trials get larger and longer in the past two decades, numerous group sequential designs have been developed to ensure overall type I and power requirements. Among them, Pocock (1977), O'Brien and Fleming (1979) and Wang and Tsiatis (1987) are three of the well-known ones. Non-binding upper efficacy boundaries, by definition, are defined without considering stopping for futility lower boundaries, which allow analysis of overrunning data when efficacy boundary was already crossed and efficacy was claimed in previous stage. Hence one-sided asymmetric group sequential designs with non-binding upper efficacy boundaries are considered in this paper. Group sequential trials for, to plan the duration of group sequential trials for survival response, Kim and Tsiatis (1990) provided algorithm to calculate the required length of the period for continued observation in the group sequential setting when the accrual period length is fixed under the scenario that there is no censoring process competing with time to failure. Different from Kim and Tsiatis (1990), we allow to have time to censoring process; and we search for the length of accrual period instead of searching for the length of the period of continued observation as we deem that, in real clinical practice, randomized subjects should have to expose to the study medication for a period long enough to evaluate drug effect and, hence, length of accrual is calculated according to a fixed length of the period of the continued observation. A required period of continued observation for every subject in the trial allows biological systems to respond to the investigational drug so that the trial results on treatment effect are more clinically interpretable.

Section 4.2 lays out the notations and other preliminaries for fixed sample design with survival response and then for group sequential designs. Section 4.3 describes the calculation of accrual period length, accumulated number of patients and real times for group sequential analyses with a period of continued observation after accrual. Section 4.4 lays out the overall characteristics for such group sequential designs for a wild range of cases. Section 4.5 discusses results and potential usage of proposed method in practice as compared with common survival clinical trials without a period of continued observation.

Section 4.2: Preliminaries

There is an accrual period of s_a years, during which patients are uniformly randomized into either the treatment group or the placebo group with ratio of $A{:}\,1$. After all qualified patients are randomized, there is a period called continued observation, during which all subjects remain treated in the randomization phase for another s_f years. Time to failure for control subjects is exponentially distributed with constant hazard rate λ_c, hence with median time $M_c = \ln(2)/\lambda_c$. To test against the null hypothesis of equal survival, i.e.,$\ln(\Delta) = 0$, where $\Delta = \frac{\lambda_c}{\lambda_E}$, λ_E being the hazard rate for experimental group subjects, we wish to have a pre-specified power against one-sided alternative of $\ln(\Delta) > 0$, or $\Delta > 1$. During the randomization phase, time to failure are independently and identically distributed (referred to as ‘i.i.d.’) within groups and independent of entry time as well as being independent of time to censoring process, where time to censoring are i.i.d.s with exp (ϕ), with the same hazard rate ϕ in both groups. The reason to use $\ln(\widehat{\Delta})$ instead of $\widehat{\Delta}$ is because $\ln(\widehat{\Delta})$ is less skewed and has a more accurate asymptotic approximation, where $\widehat{\Delta}$ is the estimated hazard ratio.

For a fixed sample design, to test $H_0\colon \ln(\Delta) = 0$ vs. $H_A\colon \ln(\Delta) > 0$ at one-sided significance level of $\alpha/2$ and power of $1-\beta$ under alternative hypothesis, we need to link log hazard ratio with the overall type I and II error requirements using asymptotic properties of the logrank statistic; and then calculate accrual period length to ensure required number of events, which is closely associated with testing power. In Appendix of Rubinstein, Gail and Santer (1981) proved that $\ln(\widehat{\Delta})$ is asymptotically normally distributed with mean $\ln(\Delta)$ and variance $\sigma^2 = [E(e_c)]^{-1} + [E(e_E)]^{-1}$, where $E(e_c)$ and $E(e_E)$ are expected number of events accumulated at the end of the trial for control and experimental groups respectively and the total trial duration is $s_a + s_f$. Of note, symbol A in the following equations is the randomization ratio of treatment group relative to placebo group, where $A = 1$ is used for all examples in this paper to indicate equal randomization in the randomization phase.

From Appendix 1B' at end of this paper, when accrual rate is m per year, we have:

$$E(e_C(s)) = \frac{m\,\lambda_C}{(A+1)\,(\lambda_C+\phi)}\left[s_a - \frac{\exp[-(\lambda_C+\phi)s_f]-\exp[-(\lambda_C+\phi)(s_a+s_f)]}{\lambda_C+\phi}\right] \quad \text{and}$$

$$E(e_E(s)) = \frac{m\,A\lambda_E}{(A+1)\,(\lambda_E+\phi)}\left[s_a - \frac{\exp[-(\lambda_E+\phi)s_f]-\exp[-(\lambda_E+\phi)(s_a+s_f)]}{\lambda_E+\phi}\right]$$

Because $\ln(\Delta) = 0$ under the null hypothesis, the asymptotic one-sided size $\alpha/2$ test of $H_0\ vs.H_A$ rejecting null for $\ln(\widehat{\Delta}) > \hat{\sigma}\ Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the standard normal $(1-\alpha/2)$ quantile and Z is the standard normal random variable. To have power $1-\beta$, we then have to have $P_{H_A}(\ln(\widehat{\Delta}) > \hat{\sigma}\ Z_{1-\alpha/2}) = 1-\beta$. Using normal distribution property, we obtain

$$[E(e_c)]^{-1} + [E(e_E)]^{-1} = [\frac{\ln(\Delta)}{(Z_{1-\alpha\backslash 2}+Z_{1-\beta})}]^2 \qquad (4.1)$$

Moving the right-hand side of Equation 1 to its left side, a function equal to zero (i.e., f(s_a)=0) is created. Utilizing Newton-Raphson method, we can reversely find accrual time of s_a for the fixed sample design. Derivative of $f(s_a)$ contains two components: $\frac{dE(e_c)^{-1}}{ds_a}$ and $\frac{dE(e_E)^{-1}}{ds_a}$, which are derived in Appendix 4.1B'.

Additionally, if under null hypothesis, when $E(e_c) = E(e_E) = \frac{d_{fix}}{2}$, with d_{fix} being the total number of events accumulated at the end of the trial for a fixed sample design, variance of log hazard ratio $\sigma_{fix}^2 = [E(e_c)]^{-1} + [E(e_E)]^{-1} = \frac{4}{d_{fix}}$. The standardized test statistic based on estimate of log hazard ratio is asymptotically equal to logrank statistic. That is $\frac{\ln(\hat{\Delta})}{\hat{\sigma}} = Z$.

We now explore the relationship between $\ln(\widehat{\Delta})$ and the logrank test statistic in a group sequential setting. Since the sequential version of Logrank test statistic $T(s) = \ln(\widehat{\Delta}) * \frac{1}{\hat{\sigma}^2}$ $= \frac{1}{\hat{\sigma}} Z$, where $T(s)$ has asymptotical normal distribution of $(s) \sim N(\ln(\Delta)\,\mathrm{V(s)},\ \mathrm{V(s)})$, $\mathrm{V(s)}$ is the reciprocal of the variance of $\ln(\widehat{\Delta})$ at time s (or called as the Fisher's information for $\ln(\widehat{\Delta})$ at time s, with $s \in (0,\ s_a +\ s_f)$), which is approximately $\frac{d(s)}{4}$, or precisely $V(s) = \frac{1}{[E(e_c(s))]^{-1}+[E(e_E(s))]^{-1}}$, when $s = s_a +\ s_f$. Z, as before, is the standard normal random variable.

Normal approximation of the sequential Logrank was first proposed by Armitage (1975), verified via simulation by Gail, DeMets, and Slud (1981), refined by Jennison and Turnbull (1984), and finally proved by Tsiatis (1982), Sellke and Siegmund (1983), and Slud (1984).

To implement a particular group sequential test, Fisher's information for a group sequential trial is obtained by multiplying the Fisher's information of fixed sample design by a factor (denoted as $1/R_{gsd}$) to ensure power of testing the null against the alternative in the group sequential setting (Jennison and Turnbull, 2002). Therefore, the variance of sequential test at time t_i is the time fraction multiplying R_{gsd}, and then multiplying variance of the corresponding fixed sample design. Suppose analysis time s becomes s_i, $\mathrm{i} = 1, \dots, \mathrm{K}$, where t_i is the information fraction used at s_i, and K analyses are performed for a group sequential design, variance at s_i

$$V(s_i) = t_i * R_{gsd} * \sigma_{fix}^2 = \frac{t_i * R_{gsd} * d_{fix}}{4} \tag{4.2}$$

Alternatively, we can calculate variance of $\ln(\widehat{\Delta})$ at time s_i as

$$V(s_i) = E(e_c(s_i))]^{-1} + [E(e_E(s_i))]^{-1} \tag{4.3}$$

Equating Equation 4.2 with Equation 4.3, we can easily find a way to search real time for interim analysis at time s_i (see Appendices 4.1A and 4.1B), as all numbers in the right hand of Equation 4.2 are given by design parameters and s can be searched using Newton-Raphson algorithm. Given a function f defined over s_i, and its derivative f', we begin with a first guess $\mathrm{s}_{\mathrm{i},0}$ for a root of the function f. Provided the function satisfies all the assumptions made in the derivation of the formula, a better approximation $\mathrm{s}_{\mathrm{i},1}$ is $\mathrm{s}_{\mathrm{i},1} = \mathrm{s}_{\mathrm{i},0} - \frac{\mathrm{f}(\mathrm{s}_{\mathrm{i},0})}{\mathrm{f}\prime(\mathrm{s}_{\mathrm{i},0})}$. The process is repeated as $s_{i,n+1} = s_{i,n} - \frac{f(s_{i,n})}{f\prime(s_{i,n})}$ until a sufficiently accurate value s_i is reached.

That is, target function f is as follows:

$\frac{4}{t_i * R_{gsd} * d_{fix}} - [E(e_c(s_i))]^{-1} + [E(e_E(s_i))]^{-1} = 0$. Based on Appendix 4.1A, when $s_i \le s_a$,

$E(e_c(s_i)) = \frac{m\lambda_C}{(A+1)(\lambda_C+\phi)}[s_i - \frac{1-\exp[-(\lambda_C+\phi)s_i]}{\lambda_C+\phi}]$ and

$E(e_E(s_i)) = \frac{mA\lambda_E}{(A+1)(\lambda_E+\phi)}[s_i - \frac{1-\exp[-(\lambda_E+\phi)s_i]}{\lambda_E+\phi}]$.

When $s_i > s_a$, $E(e_c(s_i)) = \frac{m\lambda_C}{(A+1)(\lambda_C+\phi)}[s_a - \frac{\exp[-(\lambda_C+\phi)(s_i-s_a)]-\exp[-(\lambda_C+\phi)s_i]}{\lambda_C+\phi}]$ and

$E(e_E(s_i)) = \frac{mA\lambda_E}{(A+1)(\lambda_E+\phi)}[s_a - \frac{\exp[-(\lambda_E+\phi)(s_i-s_a)]-\exp[-(\lambda_E+\phi)s_i]}{\lambda_E+\phi}]$.

Searching for s using Newton-Raphson needs $f'(s_i)$, which involves $\frac{dE(e_c(s_i))^{-1}}{ds}$ and $\frac{dE(e_E(s_i))^{-1}}{ds}$ Both are provided in Appendices 4.1A and 4.1B for $s_i < s_a$ and $s_i > s_a$, respectively.

Section 4.3: Design of Group Sequential Trials with a Period of Continued Observation

For a group sequential design, to test $H_0\colon \ln(\Delta) = 0$ vs. $H_A\colon \ln(\Delta) > 0$ with $i = 1,2, \dots K,$ we have to satisfy both type I and II error requirements under group sequential settings. Considering a group sequential trial with K planned analyses, let θ be the parameter of interest, a measure of placebo-drug difference and assume it can be estimated from trial data. The distribution of statistics Z_1 , Z_2 , …, Z_K are derived from cumulative data up to stages from 1,2 … $K,$ and it follows a canonical joint form (Chapter 3, Jennison and Turnbull, 2000) of multivariate normal distribution with E(Z_i) =$\theta\sqrt{t_i}$ and Cov(Z_i, Z_j)=$\sqrt{t_i/t_j}$, 1≤ $i \le j \le K$ and $\{t_1, \dots, t_K\}$ are information levels for parameter θ, whith final $t_K = 1$.

Startng with notations in Section 4.2, where time s is on continuous scale ranging from 0 to end of study time $s_a + s_f$, analysis times in group sequential design are discretized at K time points. Now, analysis time s becomes $s_i, \mathrm{i} = 1, \dots, \mathrm{K}$, where $s_K = s_a + s_f$. Accordingly, to accommodate group sequential notations, we denote, on the discretized time points instead, $e_{c,i}$ is the accumulative number of events at Stage i, which is the same as $e_c(s)$ in Section 4.2, with $s = s_i$. Simliarly, $e_{E,i}, d_i, V_{i,}\mathrm{i} = 1, \dots, \mathrm{K},$ are discretized versions of $e_E(s), d(s)$ and $V(s)$ respectively with $s = s_i$.

Because of asymptoticl normality of $T(s)$ (with $s = t_i$) mentioned in Section 4.2, standardized logrank statistic at (Chapter 13.2, Jennison and Turnbull),

$\hat{\theta} = \frac{\ln(\hat{\Delta})}{\hat{\sigma}}$ = Z obtained at Stage i aproximately has the canonical joint distribution,

withstandardized information level of

$$t_i = \frac{V_i}{V_K} = ([E(e_{c,i})]^{-1} + [E(e_{E,i})]^{-1})/([E(e_{c,K})]^{-1} + [E(e_{E,K})]^{-1}) \approx (\frac{4}{d_i})/(\frac{4}{d_K})$$

For a group sequential test, upper efficacy boundaries (Equation 4.4) are made to preserve type I error under null hypothesis. Non-binding upper boundaries $\{u_1, \ldots, u_K\}$ are used as their calculations do not depend on lower bounds of $\{l_1, \ldots, l_K\}$. Fisher's information vector, which is $R_{gsd} * \{t_1, \ldots, t_K\}$ and a multiple of standardized information vector, together with Kim-DeMets (1987), is used to search for the lower boundaries to maintain per-specified power under alternative hypothesis (Equation 4.5).

$$P_{H_0}\{Z_1 \geq u_1 \cup Z_2 \geq u_2 \cup \cdots \cup Z_K \geq u_K\} = \frac{\alpha}{2} \tag{4.4}$$

$$P_{H_A}\{Z_1 \geq u_1\} + P_{H_A}\{l_1 \leq Z_1 \leq u_1, Z_2 \geq u_2\} + \cdots + P_{H_A}\{l_1 \leq Z_1 \leq u_1, \ldots, l_{K-1} \leq Z_{K-1} \leq u_{K-1,} Z_K \geq u_K\} = 1 - \beta \tag{4.5}$$

Tables and Figures in this paper are created using O'Brien and Fleming (1979), Pocock (1977) and Wang and Tsiatis (1987) with shape parameter of 0.15 as efficacy upper boundaries respectively. For lower bounds $\{l_1, \ldots, l_K\}$, power spending is used with shape parameter of 0.8. That is: $f(t_i, \beta) = \beta * t_i^{0.8}, i = 1,2, \ldots, K$. For a equally-spaced three-stage group sequential design (i.e., $t^{(1)} = (0.33, 0.67, 1)$), the cumulative type II error when overall $\beta = 0.2$ is $f(t, \beta) = (0.082, 0.145, 0.2)$.

Here are the steps to calculate design parameters for group sequential trials for survival response:

1) Use α, β and log hazard ratio under alternative hypothesis to calculate required number of events for fixed sample design d_{fix}.
2) Use Equations 4.4 and 4.5 to calculate $\{l_1, \ldots, l_K\}$, $\{u_1, \ldots, u_K\}$, and R_{gsd}.
3) Given s_f and $t_K = 1$, search for s_a for a group sequential design to ensure power of group sequential test by obtaining $d_{fix} * R_{gsd}$. And the second derivatives of target function f used in Newton-Raphson search are provided in Appendix 4.1B'.

4) For the ith interim analysis, inverse search of real time s_i $i = 1, \ldots, K-1$, for the ith interim analysis is performed using Newton-Raphson algorithm as explained in Section 4.2 with the second derivative of target function f provided in Appendices 4.1A and 4.1B for $s_i \leq s_a$ and $s_i > s_a$, respectively. Of note, the searching process can start from initial real time vector $s_{i,0} = \ (s_a + s_f) * t_i$.

5) Number of patients recruited at stage $i, i = 1, \ldots, K,$ is $N_i = ms_i$ if $s_i \leq s_a$, otherwise $N_i = ms_a$ if $s_i > s_a$.

Section 4.4: Examples

With all examples with one-sided type I error of 0.025 and power of 0.8 , K=3 three-stage group sequential designs, median time of failure for the control group = 1 year, three different information times are chosen: $t^{(1)} = (0.33, 0.67, 1)$, $t^{(2)} = (0.5, 0.75, 1)$, and $t^{(3)} = (0.2, 0.8, 1)$ to represent equal increment of time fraction, interims occurring in the later part of the study and first interim occurred in the early part and later ones in the later part for $t^{(1)}, t^{(2)}$ and $t^{(3)}$, respectively. Hazard rate of λ_c/λ_E is ranging from 1.3 to 3 in Figures 4.1-4.2. Lower rate of accrual with $m = 50$ per year is used to compare with brisk accrual of $m = 240$ per year which is 20 patients per month. O'Brien and Fleming (referred to as 'OBF'), Pocock and Wang and Tsiatis(referred to as 'WT') are plotted in red, blue and green respectively in Figures 4.1- 4.2. 'Fixed' denotes cases for fixed sample design.

For Figures 4.1- 4.2 as well as Tables 4.4 - 4.6, there are three types of design features in terms of with/without censoring and with/without a period of continued observation. Types A, B and C are depicted using solid line, dotted line and dashed line respectively in Figures 4.1- 4.2.

Type A: With censoring ($\phi = \lambda_c/2$) and no continued observation ($s_f = 0$)

Type B: No censoring($\phi = 0$) and no continued observation ($s_f = 0$)

Type C: No censoring($\phi = 0$) and continued observation for $s_f = 1$ year

Comparing Type B with Type A shows the impact of competitive censoring on enlarging necessary accrual time and trial duration and comparing Type C against Type B gives the effect of adding a continued observation period on shortening accrual time but enlarging total trial duration. Varying hazard ratios and slow accrual versus quick enrolment rate on the extent of the above are assessed by evaluating Types A, B and C under a certain combination of hazard ratio and accrual rate.

Table 4.1 shows that eliminating censoring decreases required accrual time more for low accrual rate than that of high accrual rate: under $t^{(1)}$**,** by 4.57 years for OBF with rate of 50 per year and hazard ratio of 1.3 (from 15.18 to 10.61), while only 0.67 years (from 3.98 to 3.31) for rate of 240 per year at the same low hazard ratio 1.3; similarly but to a much lesser extent for high hazard ratio of 3: by 2.10 years (from 2.39 to 2.10) for m=50 per year as compared with by 0.05 year (from 0.98 to 0.93) for m=240 per year. Similar trends exist in all three group sequential designs and all three time information vectors. This confirms that the power of detecting treatment difference for survival trials only depends on number of events. When accrual rate is low and/or hazard ratio is small, more time is needed to accumulate events to ensure power. Therefore, the impact of competing from censoring will enlarge the accrual time more for either lower accrual rate and/or lower hazard ratio as events will take longer time to occur in the treatment group. Table 4.1 also shows that including one year of continued observation always shortens required accrual years: from 10.61 to 9.86 years, from 2.10 to 1.36 years, from 3.31 to 2.59 years and from 0.93 to 0.38 years for OBF tests performed at $\mathrm{t}^{(1)}$ information times with m=50 per year and $\Delta = 1.3$, m=50 per year and $\Delta = 3.0$, m=240 and $\Delta = 1.3$ and m=240 per year and $\Delta = 3.0$ respectively, where the saving for the last case with both high accrual rate and high hazard ratio is more than 50%!

Table 13(Tab. 4.1): Accrual time for group sequential designs for low or high hazard ratio

Table 4.1: **Accrual time for group sequential designs for low or high hazard ratio (1.3 vs. 3.0) and slow or brisk accrual rate (50 per year vs. 240 per year), $\alpha/2$=0.025, β=0.2, $\phi = \lambda_c/2$ for Type A and $s_f = 1$ years for Type C.**

		Fixed			OBF			Pocock			WT		
		A	B	C	A	B	C	A	B	C	A	B	C
a= 50 Δ = 1.3	$t^{(1)}$	**15.42**	**10.78**	**10.03**	**15.18**	**10.61**	**9.86**	**16.77**	**11.62**	**10.87**	**18.72**	**12.85**	**12.10**
	$t^{(2)}$				**15.33**	**10.70**	**9.95**	**16.18**	**11.24**	**10.49**	**18.89**	**12.96**	**12.21**
	$t^{(3)}$				**15.18**	**10.61**	**9.86**	**16.90**	**11.70**	**10.95**	**18.65**	**12.81**	**12.06**
a = 50 Δ = 3.0	$t^{(1)}$	2.16	1.93	1.23	**2.39**	**2.10**	**1.36**	**2.55**	**2.23**	**1.48**	**2.75**	**2.38**	**1.62**
	$t^{(2)}$				**2.40**	**2.11**	**1.37**	**2.49**	**2.18**	**1.43**	**2.77**	**2.40**	**1.63**
	$t^{(3)}$				**2.38**	**2.10**	**1.36**	**2.57**	**2.24**	**1.48**	**2.75**	**2.38**	**1.61**
a = 240 Δ = 1.3	$t^{(1)}$	4.02	3.34	2.62	**3.98**	**3.31**	**2.59**	**4.32**	**3.55**	**2.82**	**4.73**	**3.84**	**3.11**
	$t^{(2)}$				**4.01**	**3.33**	**2.61**	**4.19**	**3.46**	**2.74**	**4.77**	**3.86**	**3.13**
	$t^{(3)}$				**3.98**	**3.31**	**2.59**	**4.35**	**3.57**	**2.84**	**4.72**	**3.83**	**3.10**
a = 240 Δ = 3.0	$t^{(1)}$	0.87	0.83	0.33	**0.98**	**0.93**	**0.38**	**1.04**	**0.98**	**0.41**	**1.11**	**1.04**	**0.46**
	$t^{(2)}$				**0.98**	**0.93**	**0.38**	**1.02**	**0.96**	**0.40**	**1.11**	**1.05**	**0.46**
	$t^{(3)}$				**0.98**	**0.93**	**0.38**	**1.04**	**0.98**	**0.42**	**1.11**	**1.04**	**0.46**

Figures 4.1- 4.2 are the counterparts of Figure 1 in Rubinstein, et al., (1981), but expanded to include group sequential designs. Accrual time s_a required to conduct a test against $H_0\colon \ln(\Delta) = 0$ is plotted on the x- axis with size $\alpha/2 = 0.025$ and power of 0.8 ($\beta = 0.2$) to detect the alternative Δ on the y-axis. For all curves in Figures 4.1- 4.2, median time to failure for control group subjects is always 1 year. Figure 4.1 plots the curves for long duration trials with slow accrual (m=50 per year) while Figure 4.2 plots short duration with a brisk accrual

(m=240 per year). Within each set (one particular design with a certain information time vector), consisting of three types, the upper curve represents Type A, the case with censoring present ($\phi = \lambda_c/2$ and $s_f = 0$); the middle curve represents Type B, the case with no censoring and no continued observation period ($\phi = 0$ and $s_f = 0$); and the lower curve represents Type C, the case with one-year of continued observation period after accrual ends ($\phi = 0$ and $s_f = 1$). Figures 4.1 - 4.2 and Tables 4.2 - 4.4 show that, similar to fixed sample designs, in group sequential designs, eliminating one-year of continued observation only reduces $1/4$ year in total trial duration (from 14.25 years to 14 years for OBF, $t^{(1)}$, $\Delta = 1.25, \mathrm{m} = 50$ per year), that is to say, accrual time increases for $3/4$ years. This is kind of counter-intuitive but quite inspiring: there are indeed two ways to collect events for a survival trial, recruiting more patients or following patients in the trial for a longer time. An ideal way needs to be identified, on one hand, to account for disease characteristics for enough exposure so that treatment effect can take place; and on the other hand to shorten time length and meet economic cost limitations. Half of a year saving in time or fifty less subjects to be recruited matters a lot in today's drug development process in face of harsh competition and high cost in conducting clinical trials. Eliminating one-year of continued observation reduces very little for a short duration trial with a rapid accrual, i.e., m = 240 per year, from 4.35 years to 4.07 year for OBF, $t^{(\mathbf{1})}$, $\Delta = 1.25$; in other words, only increases in accrual time by 0.72 years. Subsequently, this elimination will result in accrual of a large chunk of patients to compensate for lacking a continued observation period.

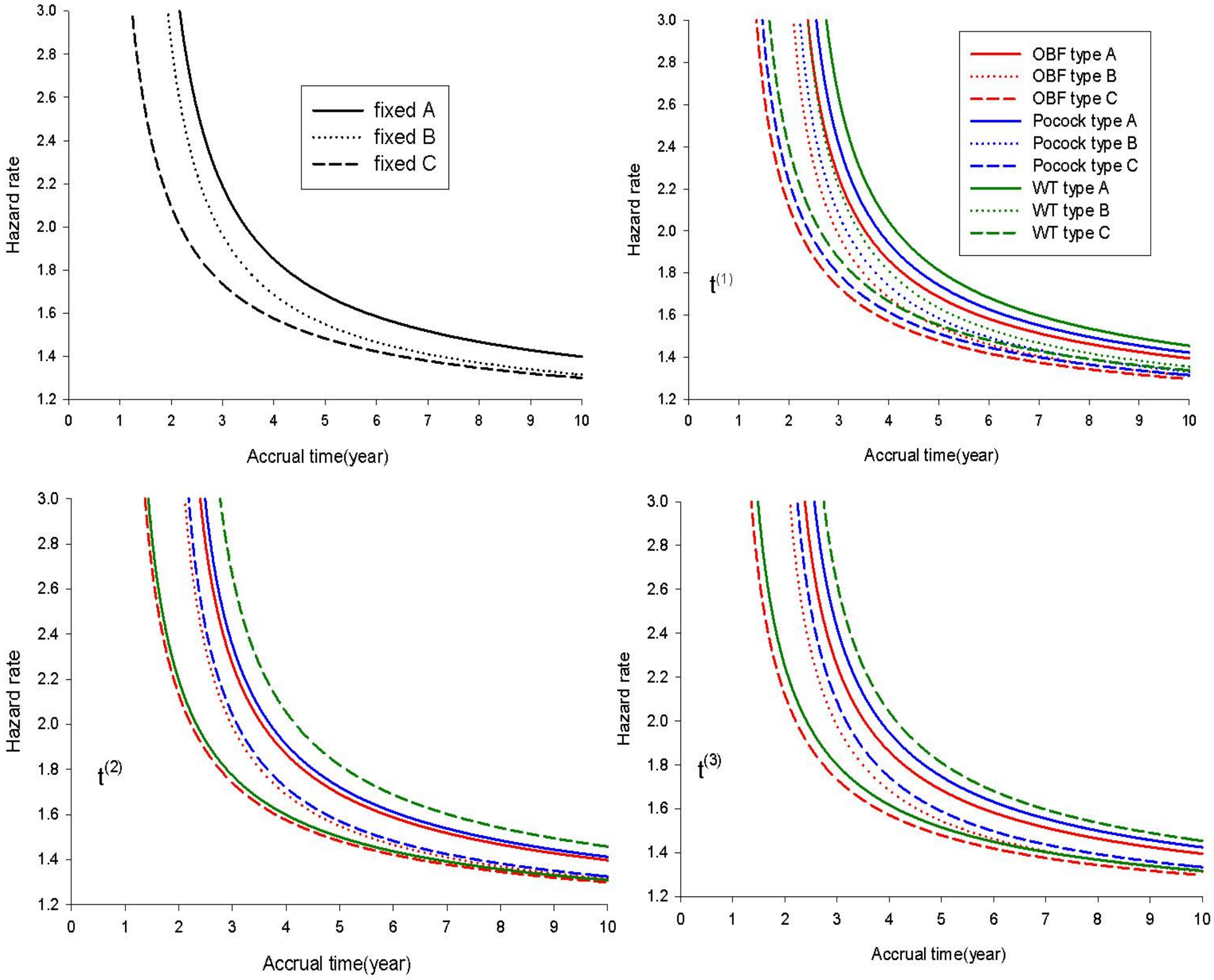


Figure 12(Fig. 4.1): Required accrual time (slow) vs. hazard ratio

Figure 4.1: Required accrual time vs. hazard ratio (from 1.3 to 3.0) for accrual rate of 50 per year, $\alpha/2$=0.025, and β=0.2 (color figure available online).

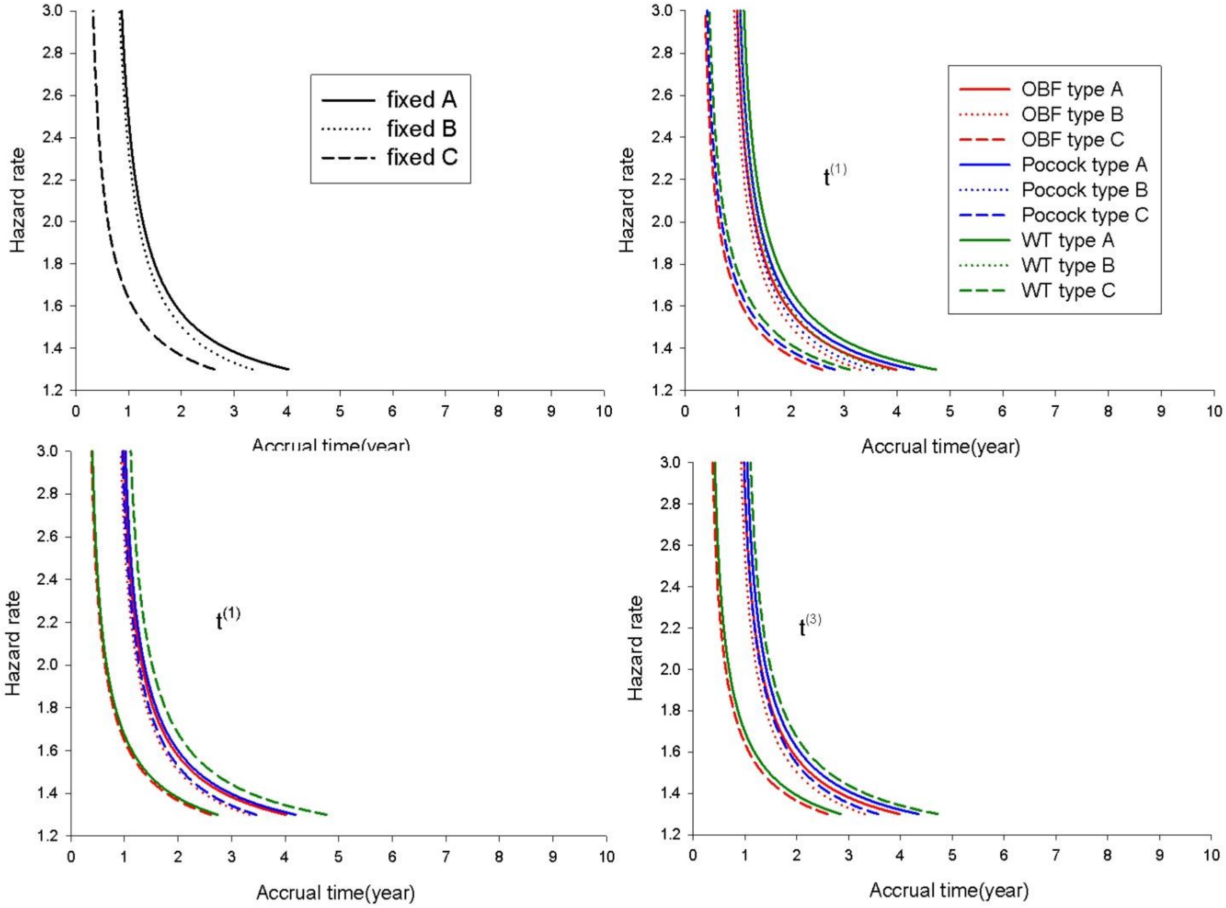


Figure 13(Fig. 4.2): Required accrual time (fast) vs. hazard ratio

Figure 4.2: Required accrual time vs. hazard ratio (from 1.3 to 3.0) for accrual rate of 240 per year, $\alpha/2$=0.025, and β=0.2 (color figure available online).

Tables 4.2 - 4.4 furthermore show that, in contrast to a long trial with slow accrual (m = 50 per year), for short trial with rapid accrual rate (i.e., m = 240 per year), adding censoring process will increase accrual time, subsequently in total trial time to a less extent. Let's take OBF, $t^{(1)}$, $\Delta = 1.25, \mathrm{m} = 240$ per year $\mathrm{s_f}$ = 0 as an example, censoring ($\emptyset = 0.5\lambda_c$) adds 1 years in total trial duration (from 4.07 years to 5.07 years) while for 6.4 years (from 14 years to 20.40 years) when with a shorter trial associated with low accrual time of m = 50. Actually, from Figures 4.1- 4.2, we can also see adding censoring changes little in accrual time for long trials with brisk accrual unless hazard ratio is less than 2. On the other hand, this reminds us that accounting for censoring in designing group sequential survival trials are important when we have a long trial associated with slow accrual and/or alternative hazard ratio is small. In such cases, ignoring censoring will result in underestimated trial accrual time and total trial duration, which leads to inadequate design preparation. Unfortunately, ignoring censoring widely exists in designing clinical trials with survival endpoint from practices nowadays.

Table 14(Tab. 4.2): Total trial duration for OBF group sequential trials

Table 4.2: Total trial duration for OBF group sequential trials when information vector is $\mathbf{t^{(1)}}$, $\boldsymbol{m}$ = 50 per year or 240 per year, $\mathbf{s_f}$ = 0, 0.5, 1, or 2 years, $\boldsymbol{\frac{\alpha}{2}}$= 0.025, and $\boldsymbol{\beta}$ = 0.2.

	Δ	$\emptyset$=0		$\emptyset=0.25\lambda_c$		$\emptyset=0.5\lambda_c$		$\emptyset=\lambda_c$	
		50	240	50	240	50	240	50	240
s_f=0	1.25	14.00	4.07	17.12	4.54	20.40	5.07	27.13	6.26
	1.5	5.49	2.01	6.32	2.13	7.26	2.26	9.35	2.56
	2	2.96	1.22	3.22	1.27	3.53	1.32	4.25	1.42
	3	2.10	0.93	2.23	0.95	2.39	0.98	2.74	1.04
s_f=0.5	1.25	14.06	4.15	17.21	4.63	20.50	5.17	27.27	6.39
	1.5	5.56	2.10	6.40	2.23	7.35	2.37	9.48	2.69
	2	3.03	1.34	3.30	1.39	3.62	1.44	4.36	1.56
	3	2.17	1.06	2.31	1.09	2.47	1.12	2.84	1.19
s_f=1	1.25	14.25	4.35	17.43	4.86	20.75	5.43	27.57	6.70
	1.5	5.73	2.33	6.61	2.48	7.59	2.64	9.77	3.00
	2	3.21	1.63	3.50	1.69	3.84	1.75	4.64	1.89
	3	2.36	1.38	2.51	1.41	2.69	1.45	3.10	1.53
s_f=2	1.25	14.84	4.98	18.12	5.55	21.51	6.18	28.42	7.55

	1.5	6.31	3.05	7.26	3.23	8.32	3.42	10.60	3.84
	2	3.81	2.43	4.15	2.51	4.54	2.58	5.44	2.75
	3	2.99	2.23	3.17	2.27	3.38	2.32	3.87	2.41

Table 15(Tab. 4.3): Total trial duration for Pocock group sequential

Table 4.3: Total trial duration for Pocock group sequential trials when information vector is $\mathbf{t}^{(1)}$, m = 50 or 240 per year, s_f= 0, 0.5, 1, or 2 years, $\frac{\alpha}{2}$= 0.025, and β = 0.2.

	Δ	Ø=0		Ø=0.25λ_c		Ø=0.5λ_c		Ø=λ_c	
		50	240	50	240	50	240	50	240
s_f=0	1.25	15.38	4.39	18.91	4.92	22.57	5.53	30.10	6.88
	1.5	5.93	2.14	6.88	2.27	7.95	2.42	10.31	2.77
	2	3.16	1.30	3.46	1.35	3.81	1.40	4.63	1.52
	3	2.23	0.98	2.38	1.01	2.55	1.04	2.96	1.10
s_f=0.5	1.25	15.45	4.46	19.00	5.01	22.67	5.63	30.23	7.01
	1.5	6.00	2.23	6.96	2.37	8.04	2.53	10.44	2.90
	2	3.22	1.41	3.54	1.46	3.89	1.53	4.75	1.66
	3	2.29	1.11	2.45	1.14	2.63	1.17	3.06	1.25
s_f=1	1.25	15.64	4.66	19.21	5.24	22.93	5.88	30.54	7.32
	1.5	6.17	2.46	7.17	2.62	8.28	2.80	10.73	3.20
	2	3.40	1.69	3.73	1.75	4.12	1.82	5.02	1.98
	3	2.48	1.41	2.65	1.45	2.84	1.49	3.31	1.58
s_f=2	1.25	16.23	5.28	19.90	5.93	23.68	6.64	31.38	8.17
	1.5	6.74	3.16	7.82	3.35	9.01	3.57	11.56	4.04
	2	3.99	2.48	4.37	2.56	4.81	2.64	5.82	2.84
	3	3.09	2.26	3.29	2.30	3.52	2.35	4.07	2.46

Table 16(Tab. 4.4): Total trial duration for Wang-Tsiatis (shape = 0.15) group sequential trials

Table 4.4: Total trial duration for Wang-Tsiatis (shape = 0.15) group sequential trials when information vector is $\mathbf{t}^{(1)}$, m = 50 per year or 240 per year, s_f = 0, 0.5, 1, or 2 years, $\frac{\alpha}{2}$ = 0.025, and β = 0.2.

	Δ	Ø=0		Ø=0.25λ_c		Ø=0.5λ_c		Ø=λ_c	
		50	240	50	240	50	240	50	240
s_f=0	1.25	17.09	4.77	21.10	5.39	25.24	6.09	33.74	7.64
	1.5	6.47	2.29	7.57	2.45	8.80	2.62	11.49	3.03
	2	3.40	1.38	3.75	1.44	4.15	1.50	5.10	1.64
	3	2.38	1.04	2.56	1.07	2.75	1.11	3.22	1.18
s_f=0.5	1.25	17.16	4.84	21.18	5.48	25.35	6.19	33.87	7.77
	1.5	6.54	2.38	7.65	2.54	8.89	2.73	11.62	3.15
	2	3.46	1.49	3.82	1.55	4.23	1.62	5.22	1.77
	3	2.44	1.16	2.62	1.20	2.83	1.24	3.32	1.32
s_f=1	1.25	17.34	5.03	21.40	5.70	25.60	6.44	34.18	8.08
	1.5	6.71	2.60	7.85	2.78	9.13	2.99	11.91	3.45
	2	3.63	1.76	4.01	1.83	4.45	1.91	5.49	2.09

	3	2.62	1.46	2.81	1.50	3.03	1.55	3.57	1.65
s_f=2	1.25	17.94	5.65	22.09	6.39	26.35	7.20	35.03	8.93
	1.5	7.27	3.29	8.50	3.51	9.86	3.75	12.74	4.29
	2	4.20	2.54	4.64	2.62	5.14	2.72	6.28	2.94
	3	3.21	2.29	3.43	2.34	3.69	2.39	4.31	2.52

Based on the required number of events for a group sequential design, accrual time and total trial duration for this group sequential trial can be derived. Impacts from adding censoring and eliminating observation period are addressed above in Tables 4.1- 4.4 and Figures 4.1- 4.2. There are other aspects of group sequential design that need to be explored prior to trial start as interim analyses allowing for early stopping using accumulating data needed to conducted in contrast to fixed duration fixed sample design. These parameters are: 1) real time at interim and final analyses; 2) required number of events at each analysis; and 3) accrued number of patients at each analysis. As described in Sections 4.2 and 4.3, inverse searching using Newton-Raphson is implemented to first find real time, then accumulated number of patients is calculated to ensure required number of events at each analysis so that overall power to detect treatment effect is reached.

One moderate hazard ratio, $\Delta = 2$, is picked up to tabulate operation characteristics for OBF, Pocock and Wang-Tsiatis group sequential trials, respectively. Tables 4.5 – 4.7 list design specifics which re-emphasize the impact of censoring and continued observation on trial design. Besides new features like number of patients and real time at interim, other group sequential parameters like upper and lower bounds are also tabulated. Probability and expected information under null or alternative can be obtained easily, but not included in Tables 4.5 – 4.7 due to space limitation. From a design with equal-spaced information time for OBF, as an example, we can see eliminating one-year of continued observation has bigger impact on reducing required number of patients for a long trial with brisk accrual than that of a short trial associated with slow accrual. For m = 50 per year, the required total number of patients with $\mathrm{s_f}$ = 1 is 110 patients while requiring 148 for $\mathrm{s_f}$ = 0 for design of OBF with $t^{(1)}$**.** But adding one year of continued observation will end up saving 51% patients of subjects (from n = 294 to n = 150) for brisk accrual while only adding 0.41 years in total duration (from 1.22 years to 1.63 years). From Table 4.5 – 4.7, for m=240 per year, all group sequential designs with $\mathrm{t}^{(1)}$ and $\mathrm{t}^{(2)}$ finish required accrual prior to first interim analysis, whereas the rest of the designs finish accrual at either prior to the second analysis or at or prior to the final analysis.

Table 17(Tab. 4.5): Operation Characteristics of group sequential designs

Table 4.5: Operation Characteristics of group sequential design with OBF upper bounds and beta-spending lower bounds with shape parameter of 0.8, $\alpha/2 = 0.025$, $\beta = 0.2$, hazard ratio = 2.

# of events	Information time	bounds		Real time (year)						Number of Patients						Accrual time / follow-up time (year)					
				Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240		
		a	b	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
22	0.33	0.204386	2.976604	1.70	1.55	1.54	0.70	0.68	0.67	85	78	77	169	162	150	3.53/0	2.96/0	2.21/1	1.32/0	1.22/0	0.63/1
45	0.67	1.020234	2.08901	2.68	2.33	2.31	1.04	0.99	1.10	133	116	110	251	237	150						
67	1.0	1.709928	1.709928	3.53	2.96	3.21	1.32	1.22	1.63	176	148	110	316	294	150						
34	0.5	0.770656	2.45016	2.22	1.97	1.96	0.89	0.85	0.88	111	99	98	214	203	152	3.55/0	2.98/0	2.23/1	1.32/0	1.23/0	0.63/1
51	0.75	1.194913	2.000547	2.91	2.50	2.52	1.12	1.05	1.22	145	125	111	269	253	152						
68	1.0	1.732525	1.732525	3.55	2.978	3.23	1.32	1.23	1.63	178	149	111	318	296	152						
13	0.2	-0.35608	3.84717	1.26	1.18	1.17	0.54	0.52	0.51	63	59	58	129	125	124	3.53/0	2.96/0	2.21/1	1.32/0	1.22/0	0.63/1
54	0.8	1.390059	1.923585	3.02	2.58	2.63	1.16	1.08	1.29	151	129	110.	278	260	150						
67	1.0	1.720506	1.720506	3.53	2.96	3.21	1.32	1.22	1.63	176	148	110	316	294	150						

Table 18(Tab. 4.6): Operation Characteristics of group sequential designs

Table 4.6: Operation Characteristics of group sequential design with Pocock upper bounds and beta-spending lower bounds with shape parameter of 0.8, $\alpha/2$ = 0.025, β=0.2, hazard ratio = 2.

# of events	Information time	bounds		Real time (year)						Number of Patients						Accrual time / follow-up time (year)					
				Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240		
		a	b	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
25	0.33	0.291524	1.992737	1.82	1.65	1.64	0.75	0.72	0.71	91	82	82	179	172	165	3.81/0	3.16/0	2.40/1	1.40/0	1.30/0	0.69/1
50	0.67	1.161621	1.992735	2.88	2.48	2.46	1.11	1.04	1.15	144	124	120	266	250	165						
75	1.0	1.992734	1.992734	3.81	3.16	3.40	1.40	1.30	1.69	190	158	120	337	312	165						
36	0.5	0.82858	1.947182	2.31	2.04	2.03	0.92	0.87	0.90	115	102	102	221	210	160	3.70/0	3.09/0	2.33/1	1.37/0	1.27/0	0.66/1
54	0.75	1.285168	1.947182	3.03	2.59	2.60	1.16	1.09	1.25	151	130	116	278	261	160						
72	1.0	1.947181	1.947181	3.70	3.09	3.33	1.37	1.27	1.66	185	154	116	329	205	160						
15	0.2	-0.28265	2.002045	1.35	1.25	1.25	0.57	0.55	0.55	68	63	62	137	133	131	3.83/0	3.18/0	2.42/1	1.41/0	1.30/0	0.69/1
60	0.8	1.572608	2.002045	3.27	2.77	2.80	1.24	1.15	1.35	163	138	121	297	277	166						
76	1.0	2.002039	2.002039	3.83	3.18	3.42	1.41	1.30	1.69	191	159	121	338	313	166						

Table 19(Tab. 4.7): Operation Characteristics of group sequential designs

Table 4.7: Operation Characteristics of group sequential design with WT upper bounds (shape = 0.15) and beta-spending lower bounds with shape parameter of 0.8, $\alpha/2$ = 0.025, β = 0.2, hazard ratio = 2.

# of events	Information time	bounds		Real time (year)						Number of Patients						Accrual time / follow-up time (year)					
				Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240			Accrual rate=50			Accrual rate=240		
		a	b	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
28	0.33	0.392837	3.009054	1.96	1.76	1.75	0.80	0.76	0.75	99	88	87	191	183	181	4.15/0	3.40/0	2.63/1	1.50/0	1.38/0	0.76/1
56	0.67	1.288984	2.348463	3.12	2.66	2.64	1.19	1.11	1.20	156	133	132	285	267	182						
84	1.0	2.041314	2.041314	4.15	3.40	3.63	1.50	1.38	1.76	207	170	132	361	332	182						
43	0.5	1.002881	2.631239	2.57	2.25	2.23	1.01	0.95	0.97	129	112	112	242	229	183	4.18/0	3.42/0	2.65/1	1.51/0	1.39/0	0.76/1
64	0.75	1.479783	2.283118	3.39	2.86	2.86	1.28	1.19	1.33	170	143	133	306	285	183						
85	1.0	2.064428	2.064428	4.18	3.42	3.65	1.51	1.39	1.76	209	171	133	363	334	183						
17	0.2	-0.21179	3.59089	1.44	1.33	1.32	0.60	0.58	0.58	72	66	66	145	140	139	4.13/0	3.39/0	2.62/1	1.50/0	1.38/0	0.76/1
67	0.8	1.678852	2.210452	3.52	2.95	2.97	1.31	1.22	1.40	176	148	131	315	293	181						
84	1.0	2.044384	2.044384	4.13	3.40	3.63	1.50	1.38	1.76	207	169	131	360	332	181						

Section 4.5: Discussion

Competitive censoring is normally not considered at the stage of designing a survival trial prior to trial start. Normal practice is that: a required number of events is firstly calculated to ensure control of type I error when null hypothesis is true and enough power to detect the alternative hypothesis when investigational compound is effective; and then a rough number of required number to be recruited is reversely calculated assuming an overall probability of a subject resulting in an event in the randomization phase irrespective of treatment groups. During the trial, accrual process stops when the required number to be recruited is achieved, whereas trial may still be ongoing until we observe at least certain number of events to ensure power of detecting the treatment difference. So there is no specification of continued observation in the trial.

As shown from tables and figures in this paper, current trial practice has many shortcomings in not accounting for factors of accrual time, continued observation time and censoring process in calculating real time and required number of patients in a group sequential trial. The minimal length of continued observation period should come from clinical perspective and depends on disease characteristics, which is a necessary period for drug to be differentiated from comparator in the trial. Constrained on this minimum length, real length of continued observation time to be used in the trial could be chosen based on balance of required number of patients and total trial length. This paper provides a method of designing a group sequential trial with fixed length of continued observation in the presence of censoring with a trial without censoring as a special case of it. A way to search for real time of interim analysis with which searching formulas depending on if the real time is less or greater than trial accrual time. Figures and tables vividly display the impact of having censoring process and having continued observation on trial accrual time and total trial during under different scenarios with a particular combination of hazard ratio and accrual rate. Results from this paper also show the necessity of doing trial design in proposed way; as such impact could be substantial in certain situations. For example, only 0.25 years increase in total trial duration can reduce the required number of patients to be 50% or more, which is really worth serious consideration in face of harsh competition in today's world. Instead of adding a required continued observation after stopping of recruitment process which means last randomized subject will be followed up to a maximum time length in the randomization phase if the survival event has not occurred prior to it and then the trial will be ended, all subjects might only be allowed to stay in the randomization phase until a maximum length in the trial or having an event. This is often a concern for trials investigating treatment of a life-threatening disease and with subjects randomized into the placebo group in the randomization phase which poses a question on long term exposure of placebo on patients in the trial. Even for subjects who are randomized into the treatment group in the randomization phase, it is ethical not allowing them to be followed too long, as it is just an investigational drug with profile of efficacy and safety not well-investigated. Research on this topic is being worked on currently, but Appendix 2 shows mathematics as the basis for numerical calculations with the difference in using grid-search instead of Newton-Raphson search for $s_i \ i = 1, \ldots, K-1, K$ as being discretized in the presence of a cap for each subject's follow-up time after accrual. Although Software ADDPLAN® and Software EAST® has implemented group sequential design for survival data and SAS® has SEQDESIGN and SEQTEST procedures to deal with designs and analyses, there hasn't been any publication substantively assessing the impacts of a period of continued observation on operation characteristics of a particular design. This paper serves this purpose and the authors would like to share our R codes with audience upon request. Per authors' over ten years of experience of being a trial statistician, direct explorations using automated codes on a variety of scenarios considering trial-specific requirements prior to trial start are much more efficient than obtaining one set of design parameters only for one scenario after entering parameters in a step-by-step fashion into software windows and then repeat the whole process for every scenario, let alone software development normally lags behind practical needs and some applications are not yet implemented to fit current trial-specific issues. Even software already has all ingredients for trial design (normally not true at all), it is hard to be utilized for finding an optimal design regarding a specific cost function to be used in a survival trial; for example, an optimal design considering efficiency in terms of both time and detecting power. All concerns listed above led us writing up this paper to share with all trial statisticians; and optimal survival trials are being investigated by us.

References

Armitage P. *Sequential Medical Trials*. Oxford: Blackwell. 1975.

Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B* 1972; 34(2): 187–220.

Gail MH, DeMets D L, Slud EV. Simulation studies on increments of the two sample logrank score test for survival time data, with application to group sequential boundaries. In *Survival analysis.* IMS Lecture Notes, Monograph Series 2, R. Johson and J. Crowley (eds), 287-301. Hayward, California: Institute of mathematical Statistics. 1981.

George SL, Desu MM. planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Disease* 1974; **27**:15-29.

Harrington D. Linear Rank Tests in Survival Analysis. *Encyclopedia of Biostatistics*. Wiley Interscience. 2005.

Jennison C , Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman and Hall. 2000.

Jennison C, Turnbull BW. Repeated confidence intervals for group sequential trials. *Controlled Clinical Trials* 1984; 5:33-45.

Kim, K. and DeMets, D.L. (1987). Design and Analysis of group sequential tests based on Type I error spending rate functions. *Biometrika* 74, 149-54

Kim K, Tsiatis AA. Study duration for clinical trials with survival response and early stopping rule. *Biometrics*. 1990, 46(1): 81-92.

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; 50(3): 163–70.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35:549-56.

Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society, Series A* 1971; 135(2): 185–207.

Rubinstein LV, Gail MH, Santer TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease* 1981; 34:469-479.

Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1977; 38:153-162.

Wang SK, Tsiatis AA, Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 1987, 43:193-199.

Sellke K, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983; 70:315-326.

Slud EV. Sequential linear rank tests for two-sample censored survival data. *Annuals of Statistics* 1984; **12**:551-571.

Slud EV, Wei LJ. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* 1982, 77:862-868.

Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* 1982; 77:855-861.

Appendix 4.1: No Cap for Follow-up Time on Each Subject

Appendix 1A: $s \le s_a$

Let's set time to randomize first subject in the trial as anchor time 0 and assume time to censoring is present in the trial and independent of process of time to event. For a subject in the control group who was randomized at time u, at real time s, the time from randomization to evaluation time point is $s-u$, and thus the probability of this entry will result in an event is:

$$P[Y_c < W_c,\, Y_c < s-u] = \int_0^{s-u} \lambda_C \exp(-\lambda_C t)\exp(-\phi t)\, dt$$

$$= \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\,\lambda_C+\phi)(s-u)]\,]$$

$$E(e_c(s)|n_C) = n_C P(\text{event on control}) = n_C E[\,E[\,I(Y_c < W_c,\, Y_c < s-u)|u\,]\,]$$

$$= n_C \int_0^s P(\text{event on control|time from randomization to evaluation time being u}) g(u)du$$

$g(u)$ is the density of u. Based on uniform accrual in interval [0, s], $g(u) = \frac{1}{s}$.

$$E(e_c(s)|n_C) = n_C \int_0^s \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\,\lambda_C+\phi)(s-u)]\,]\frac{1}{s}du$$

$$= \frac{\lambda_C}{\lambda_C+\phi}[\,n_C - n_C\frac{1}{s}\,\frac{1-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]$$

With $n_C = \frac{ms}{A+1}$ and $n_E = \frac{mAs}{A+1}$,

$$E(e_c(s)) = \frac{m\lambda_C}{(A+1)(\,\lambda_C+\phi)}[\,s - \frac{1-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]$$

$$E(e_E(s)) = \frac{mA\lambda_E}{(A+1)(\,\lambda_E+\phi)}[\,s - \frac{1-\exp[-(\,\lambda_E+\phi)s]}{\lambda_E+\phi}\,]$$

$$\frac{dE(e_c(s))^{-1}}{ds} = \frac{\exp[-(\,\lambda_C+\phi)s]-1}{\frac{m\lambda_C}{(A+1)(\,\lambda_C+\phi)}[\,s-\frac{1-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]^2}$$

$$\frac{dE(e_E(s))^{-1}}{ds} = \frac{\exp[-(\,\lambda_E+\phi)s]-1}{\frac{mA\lambda_E}{(A+1)(\,\lambda_E+\phi)}[\,s-\frac{1-\exp[-(\,\lambda_E+\phi)s]}{\lambda_E+\phi}\,]^2}$$

Appendix 1B: $s > s_a$

$$E(e_c(s)|n_C) = n_C \int_0^{s_a} \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\,\lambda_C+\phi)(s-u)]\,]\frac{1}{s_a}du$$

$$= \frac{\lambda_C}{\lambda_C+\phi}[\,n_C - \frac{n_C}{s_a}\,\frac{\exp[-(\,\lambda_C+\phi)(s-s_a)]-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]$$

$$E(e_c(s)) = \frac{m\lambda_C}{(A+1)(\,\lambda_C+\phi)}[\,s_a - \frac{\exp[-(\,\lambda_C+\phi)(s-s_a)]-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]$$

$$E(e_E(s)) = \frac{mA\lambda_E}{(A+1)(\,\lambda_E+\phi)}[\,s_a - \frac{\exp[-(\,\lambda_E+\phi)(s-s_a)]-\exp[-(\,\lambda_E+\phi)s]}{\lambda_E+\phi}\,]$$

$$\frac{dE(e_c(s))^{-1}}{ds} = \frac{\exp[-(\,\lambda_C+\phi)s]-\exp[-(\,\lambda_C+\phi)(s-s_a)]}{\frac{m\lambda_C}{(A+1)(\,\lambda_C+\phi)}[\,s_a-\frac{\exp[-(\,\lambda_C+\phi)(s-s_a)]-\exp[-(\,\lambda_C+\phi)s]}{\lambda_C+\phi}\,]^2}$$

$$\frac{dE(e_E(s))^{-1}}{ds} = \frac{\exp[-(\,\lambda_E+\phi)s]-\exp[-(\,\lambda_E+\phi)(s-s_a)]}{\frac{mA\lambda_E}{(A+1)(\,\lambda_E+\phi)}[\,s_a-\frac{\exp[-(\,\lambda_E+\phi)(s-s_a)]-\exp[-(\,\lambda_E+\phi)s]}{\lambda_E+\phi}\,]^2}$$

Appendix 1B': when $s = s_a + s_f$, i.e. at the end of the trial, we will have:

$$E(e_C(s)) = \frac{m\,\lambda_C}{(A+1)\,(\,\lambda_C+\phi)}\left[s_a - \frac{\exp[-(\,\lambda_C+\phi)s_f]-\exp[-(\,\lambda_C+\phi)(s_a+s_f)]}{\lambda_C+\phi}\right]$$

$$E(e_E(s)) = \frac{m\,A\lambda_E}{(A+1)\,(\,\lambda_E+\phi)}\left[s_a - \frac{\exp[-(\,\lambda_E+\phi)s_f]-\exp[-(\,\lambda_E+\phi)(s_a+s_f)]}{\lambda_E+\phi}\right]$$

Taking derivative with respect to s_a, we then have:

$$\frac{dE(e_c(s))^{-1}}{ds_a} = \frac{\exp\left[-(\lambda_C+\phi)(s_a+s_f)\right]-1}{\frac{m\lambda_C}{(A+1)(\lambda_C+\phi)}[\, s_a-\frac{\exp\left[-(\lambda_C+\phi)s_f\right]-\exp\left[-(\lambda_C+\phi)(s_a+s_f)\right]}{\lambda_C+\phi}]^2}$$

$$\frac{dE(e_E(s))^{-1}}{ds_a} = \frac{\exp\left[-(\lambda_E+\phi)(s_a+s_f)\right]-1}{\frac{mA\lambda_E}{(A+1)(\lambda_E+\phi)}[\, s_a-\frac{\exp\left[-(\lambda_E+\phi)s_f\right]-\exp\left[-(\lambda_E+\phi)(s_a+s_f)\right]}{\lambda_E+\phi}]^2}$$

Appendix 4.2: With A Cap for Follow-up Time (τ) on Each Subject

Under Case 2A: $s \leq s_a$:

For a subject in the control group who was randomized at time u, at real time s, the time from randomization to evaluation time point is $s - u$, and thus the probability of this entry to result in an event is when every subject can stay in the trial for maximum time τ:

$$P[Y_c < W_c, Y_c < s-u, Y_c < \tau] = \int_0^{\min(s-u,\ \tau)} \lambda_C \exp(-\lambda_C t)\exp(-\phi t)\,dt$$

$$= \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\lambda_C+\phi)\ \min(s-u,\ \tau)]\,]$$

$E(e_c(s)|n_C) = n_C P(\text{event on control}) = n_C E[\,E[\,I(Y_c < W_c, Y_c < s-u, Y_c < \tau)|u\,]\,]$

$=$

$n_C \int_0^s P(event\ on\ control|time\ from\ randomization\ to\ evaluation\ time\ being\ u) g(u) du$

$g(u)$ is the density of u. Based on uniform accrual in interval [0, s], $g(u) = \frac{1}{s}$. Plugging in density of u,

$E(e_c(s)|n_C) = n_C \int_0^s \frac{1}{s}\ P[Y_c < W_c, Y_c < s-u, Y_c < \tau]du$

$= n_C \int_0^s \frac{1}{s}\ \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\lambda_C+\phi_C)\ \min(s-u,\ \tau)]\,]du$

$\therefore\ E(e_c) = \frac{1}{A+1} m s_a \int_0^s \frac{1}{s}\ \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\lambda_C+\phi)\ \min(s-u,\ \tau)]\,]du \qquad (4.1A)$

Similarly, $E(e_E) = \frac{A}{A+1} m s_a \int_0^s \frac{1}{s}\ \frac{\lambda_E}{\lambda_E+\phi}[\,1-\exp[-(\lambda_E+\phi)\ \min(s-u,\ \tau)]\,]du \qquad (4.2A)$

Under Case 2B: $s > s_a$:

$E(e_c(s)|n_C) = n_C \int_0^{s_a} \frac{1}{s_a}\ P[Y_c < W_c, Y_c < s-u, Y_c < \tau]du$

$\therefore\ E(e_c) = \frac{1}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\ \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\lambda_C+\phi)\ \min(s-u,\ \tau)]\,]du \qquad (4.1B)$

$E(e_E) = \frac{A}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\ \frac{\lambda_E}{\lambda_E+\phi}[\,1-\exp[-(\lambda_E+\phi) \min(s-u,\ \tau)]\,]du \qquad (4.2B)$

Under Case 2B', where real time $s = s_a + \tau$,

$E(e_c) = \frac{1}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\ \frac{\lambda_C}{\lambda_C+\phi}[\,1-\exp[-(\lambda_C+\phi) \min(s_a+\tau-u,\ \tau)]\,]du \qquad (4.1B')$

$E(e_E) = \frac{A}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\ \frac{\lambda_E}{\lambda_E+\phi}[\,1-\exp[-(\lambda_E+\phi)\ \min(s_a+\tau-u,\ \tau)]\,]du \qquad (4.2B')$

Chapter 5

Planning the Duration of a Survival Group Sequential Trial with a Fixed Follow-up Time for All Subjects

(accepted for publication in Jan 2016 by *Communication in Statistics: theory and method*)

Abstract: To explore the operation characteristics of survival group sequential trials with a fixed follow-up period, the accrual time and total trial duration to ensure power and type I error rate requirements are explained and investigated for hazard ratios ranging from 1.3 to 3.0, with slow or high accrual rate, and in the presence or absence of censoring. Impacts of hazard rate, accrual rate and competitive censoring on accrual time and subsequently on total trial duration are carefully illustrated. Real time for interim analyses, needed number of events and recruited number of subjects at time of interim analyses are also tabulated.

Key Words: Survival endpoint; Group sequential trial; a fixed follow-up period; Operation characteristics.

Section 5.1: Introduction and A Motivating Example

For time to event analysis, the logrank statistic was proposed by Nathan Mantel (1966) and was named by Richard and Julian Peto (1972).The logrank statistic can also be derived as the score test for the Cox Proportional Hazard model (Cox, David R, 1972) comparing survival curves between two groups. In terms of planning a survival trial, George and Desu (1974) proved that the total duration is minimized when we continue to randomize subjects into the double-blind phase until the end of the trial (i.e., no period of continued observation after accrual period). Rubinstein, Gail and Santer (1981) explored the impact of a period of continued observation on number of patients to be accrued to ensure a required statistical power and found: although total duration of the trial is increased a little as compared with that of the case with no continued observation period, accrual time could be reduced substantially as high as 50% or more after introducing a period of continued observation. Of note, both George and Desu (1974) and Rubinstein, Gail and Santer (1981) only focused on fixed sample designs.

As trials get larger and longer in the past two decades, trials are analyzed using accumulating data periodically to allow stopping early if treatment effect is shown to be large enough and/or if there is no hope to show treatment effect even when the trial lasts to the end. Numerous group sequential designs have been developed to ensure overall type I error rate and power requirements. Among them, Pocock (1977), O'Brien and Fleming (1979) and Wang and Tsiatis (1987) are three of the well-known ones. Normal approximation of the sequential logrank was first proposed by Armitage (1975), verified via simulation by Gail, Demets, and Slud(1981), refined by Jennison and Turnbull (1984), and finally proved by Tsiatis (1982), Sellke and Siegmund (1983), and Slud (1984). In group sequential trials with survival endpoints, to plan the duration of group sequential trials for survival response, Kim and Tsiatis (1990) searched required length of the period for continued observation in group sequential setting when accrual period length is fixed under the scenario that there is in the absence of censoring process competing with time to failure. Group sequential survival trials with each subject followed-up with a fixed period of time is not yet explored but frequently encountered in drug development practice as the motivating example below indicates.

Section 5.1.1: A Motivating Example

Drug A with a 1-month injection interval and has been approved by FDA. A new formulation with a 3-month administration interval (referred to as 'Drug B') is being studied for the maintenance treatment effect in subjects with recent onset of schizophrenia who have been treated for four or more months of Drug A. The primary objective of a clinical trial study is to compare the efficacy of Drug B in delaying time to first treatment failure with approved active comparator Drug A, in subjects with recent onset of schizophrenia. A randomized withdrawal trial is planned and all enrolled subjects will have an open-label phase treated with Drug A to stabilize disease status before being randomized into either Drug A group or Drug B group. Time to relapse is defined in multiple dimensions as time to first occurrence in the double-blind phase of: Psychiatric hospitalization; or suicide, deliberate self-injury or clinically significant suicidal thoughts or behavior as determined by the investigator; or change in PANSS total score or in some PANSS items (details are not described here due to non-relating to design details investigated in this paper), which, from different perspectives, shows deterioration in symptom of schizophrenia after randomization. Due to the fact that subjects in both groups will be treated with active treatments, relapse rates for subjects in either group won’t be high and thus it is not easy to accumulate relapse events in the double-blind phase. Assuming relapse rate over a year for Drug A being 30%, the primary hypothesis is to determine superiority of Drug B over Dug A on maintenance effect for having 15% less in yearly relapse rate (i.e., Drug A = 30% and Drug B = 15%). A large number of events are required to ensure 80% power to establish superiority of Drug B over Drug A. A question is now raised up: Should we conduct an event-driven trial, within which all relapse-free subjects should remain in the trial till trial termination after collecting enough number of events? By doing this, many subjects will have to stay in the trial for a very long period of time due to low event rate in both groups as well as the fact that a large number of events is required for the trial due to having relatively small treatment-placebo difference by using an active comparator. Therefore, it is hard to get consented from the patients to participate in this trial because they might end up staying in the trial for too long. Hence, together with other considerations, a reasonable follow-up period, 48 weeks, was proposed by the study team to cap the duration of each subject in the double-blind phase. It is that all subjects in the double-blind phase will be followed-up until either experiencing a relapse, or early withdrawal or up to 48 weeks, whichever date comes the earliest. A side gain from this operation is: due to the majority of subjects will be administratively censored by this fixed follow-up time (i.e., remained event-free over 48 weeks in the double-blind phase), safety parameters and secondary efficacy variables can now be reasonably assessed, because, otherwise, between-group comparisons for incidence rates of safety parameters and effects overtime of secondary endpoints make no sense when the majority of subjects have a variable length in the double-blind and one group could stay substantially longer than the other. Capping the follow-up time by 48 weeks enables the administratively censored subjects, i.e., the largest cohort among all randomized subjects, censored at 48 weeks in the double-blind phase and resulting in a comparable length of exposure in the double-blind phase within this cohort regardless of treatment groups. On the other hand, comparing a trial without any requirement on a minimum length of follow-up time could result in an un-acceptable short period for a subject to expose to the study medication upon study termination, even to the shortest of only one day. This, in some sense, violates the intent-to-treat principle because there will have a big chunk of subjects being censored at study termination right after randomization without any contribution to evaluation of between-group difference in survival curves.

Section 5.2 illustrates the trial diagram for survival trials in the absence and presence of a fixed follow-up period for each subject in Figure 5.1a and 5.1b, respectively. Rational for designing a group sequential survival trial with a fixed follow-up period for each subject is discussed in Section 5.3, together with calculating design operation characteristics. Section 5.4 shows examples explored about how adding a fixed follow-up for each subject could impact clinical trial designs. In the end, Section 5.5 includes discussions and then concludes this paper.

Section 5.2: Trial Diagram

Section 5.2.1 Survival Trials without A Fixed Follow-up Time

Figure 5.1a shows survival trials without a fixed follow-up time, which is normally done in clinical trial practice. From Figure 5.1a, we can see approximate uniform randomization accrual in $[0, s_a]$ and subjects who have remained in the trial at time s_a are all followed for additional s_f months to accumulate enough events in the trial. Vertical bar "|" on the left hand of time line denotes the timing of performing randomization procedure and then the subject enter into the double-blind phase. Circle on the right hand indicates a survival event occurred on this subject during the double-blind phase while cross symbol denotes censoring prior to study termination and triangle symbol indicates administrative censoring at trial termination. From Figure 5.1a, we have 9 events and 4 censorings by time $s_a + s_f$, including one with administrative censoring because this subject was ongoing at the time of study termination. Censorings other than administrative ones could be due to withdrawal of consent, adverse events, lost to follow-up or other reasons.

Section 5.2.2 Survival Trials with A Fixed Follow-up Time

Figure 5.1b shows the trial of interest in this paper. After being randomized into the double-blind phase, each subject will be followed-up up to a fixed length of period, for example $s_f = 0.92$ years (i.e., 48 weeks) as in the motivating example. Subjects could finish end-of –study visit due to event or censoring prior to 0.92 years follow-up time. As in Figure 5.1a, vertical bar "|" on the left hand of time line denotes date of randomization and circle indicates event times. Administrative censorings (triangle symbol) will occur due to time truncation. Note that time to administrative censoring in Figure 5.1b is fixed as of s_f years for every subject while it could be a variable number in (0, $s_a + s_f$] in Figure 5.1a. Besides, time to event in Figure 5.1b is also truncated by s_f, while being in the range of 0 to $s_a + s_f$ in trials without a follow-up time constrain as in Figure 5.1a. In Figure 5.1b, there were 5 events, 2 non-administrative censorings due to early withdrawal prior to truncation time and 6 administrative censorings due to time truncation. Time from randomization to event and censoring are both bounded by the maximum follow-up time s_f. Although it appears that the total trial duration is $s_a + s_f$ for both designs,

s_f is defined differently in two scenarios, which is the length of the continued observation period after closure of the accrual process while being the maximum follow-up time for all subjects in Figure 5.1b. When s_f is pre-defined, s_a will differ a lot in two scenarios when to detest the same alternative hypothesis and under the same conditions for accrual rate, type I error rate and power requirements.

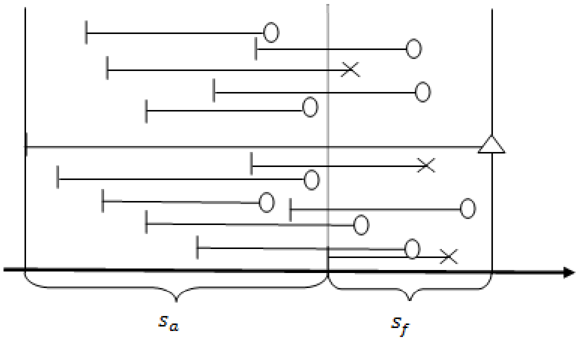


Figure 14(Fig. 5.1): Trial diagram without/with a fixed follow-up period

Figure 5.1: Trial diagram without/with a fixed follow-up period.
Figure 5.1a: Trial diagram without a fixed follow-up period. Symbol "|"denotes the timing of randomization; circle symbol indicates an event; and cross and triangle symbols denote censoring. s_a is the accrual time for the trial and s_f is the continued observation period of the trial after accrual is closed.

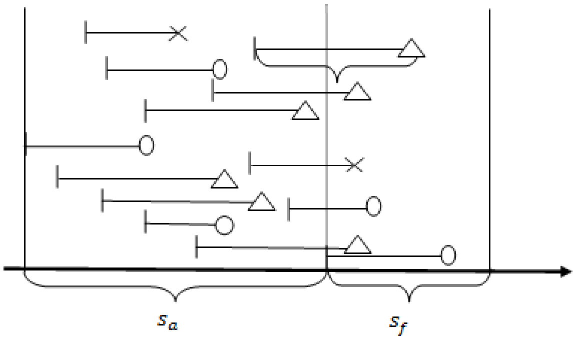


Figure 5.1b: Trial diagram with a maximum follow-up period imposed on all subjects. Symbol "|"denotes the timing of performing randomization; circle symbol indicates an event; and cross and triangle symbols denote censoring. s_a is the accrual time for the trial and s_f is the maximum follow-up time imposed on each subject .

Section 5.3: Preliminaries

Section 5.3.1: Expected Number of Events at Real Time s for Survival Trials with A Fixed Follow-up Period for All Subjects

Since patients are uniformly randomized into an interval [0, $\mathrm{s_a}$] in unit of year, the total number of subjects entering the double-blind phase N = $\mathrm{n_E}$ + $\mathrm{n_C}$ will be $\mathrm{ms_a}$ in total with recruitment rate of m per year over the $\mathrm{s_a}$ years of accrual. With randomization ratio A: 1 of treatment group ($\mathrm{n_E}$) to control group ($\mathrm{n_C}$), then expected recruitment in $\mathrm{s_a}$ years for treatment and control groups, respectively, are: $\mathrm{E[n_E]} = \frac{\mathrm{A}}{\mathrm{A+1}}\mathrm{ms_a}$ and $\mathrm{E[n_C]} = \frac{1}{\mathrm{A+1}}\mathrm{ms_a}$. Let's set time to randomize first subject in the trial as anchor time 0 and assume time to censoring is present in the trial and independent of process of time to event and accrual process. Any real time s in the trial could be either: Case A: s$\leq \mathrm{s_a}$ or Case B:$\mathrm{s} > \mathrm{s_a}$. Case B':$\mathrm{s} = \mathrm{s_a} + \mathrm{s_f}$, a special case of Case B, denotes the real time when the whole trial is terminated and the time of performing the last visit of the last patient (referred to as 'LPLV'). Assuming survival rate for treatment and control groups and censoring rate regardless of treatment assignment are exponential with rates of $\lambda_\mathrm{E}, \lambda_\mathrm{C}$ and ϕ, respectively. These three exponential random variables are mutually independent and also independent of the uniform accrual process. Let Y_i and W_i, $i = C, E$, represent random variables of time to event and time to censoring for subjects treated with control (C) and treatment (E) medications, respectively. $\mathrm{E(e_c)}$ and $\mathrm{E(e_E)}$ are expected number of events from subjects treated with control and treatment medications, respectively, accumulated up to study end, conditional upon that all subjects are followed-up up to a fixed period of s_f in the double-blind phase; and n_C and n_E are the number of subjects accrued in the control and treatment groups, respectively. Hazard ratio $\Delta = \frac{\lambda_c}{\lambda_E}$, with λ_E being the hazard rate for experimental group subjects and λ_C being the hazard rate for control-treated subjects. $\widehat{\Delta}$ is the estimated hazard ratio.

Under Case A: $s \leq s_a$:

For a subject in the control group who was randomized at time u, at real time s, the time from randomization to evaluation time point is $s - u$, and thus the probability of this entry to result in an event is:

$$P\left[Y_c < W_c,\, Y_c < s-u,\, Y_c < s_f\right] = \int_0^{\min\left(s-u,\ s_f\right)} \lambda_C \exp(-\lambda_C t)\exp(-\phi t)\, dt$$

$$= \frac{\lambda_C}{\lambda_C+\phi}\left[\,1-\exp\left[-(\lambda_C+\phi)\ \min\left(s-u,\ s_f\right)\right]\,\right]$$

$E(e_c(s)|n_C) = n_C P$(event on control)$= n_C E[\,E[\,I(Y_c < W_c,\, Y_c < s-u)|u\,]\,]$

$=$

$n_C \int_0^s P(event\ on\ control|time\ from\ randomization\ to\ evaluation\ time\ being\ u)g(u)du$

$g(u)$ is the density of u. Based on uniform accrual in interval [0, s], $g(u) = \frac{1}{s}$. Plugging in density of u,

$$E(e_c(s)|n_C) = n_C \int_0^s \frac{1}{s}\, P\left[Y_c < W_c,\, Y_c < s-u,\, Y_c < s_f\right] du$$

$$= n_C \int_0^s \frac{1}{s}\,\frac{\lambda_C}{\lambda_C+\phi}\left[\,1-\exp\left[-(\lambda_C+\phi_C)\ \min\left(s-u,\ s_f\right)\right]\,\right] du$$

$$\therefore\ E(e_c) = \frac{1}{A+1} m s_a \int_0^s \frac{1}{s}\,\frac{\lambda_C}{\lambda_C+\phi}\left[\,1-\exp\left[-(\lambda_C+\phi)\ \min\left(s-u,\ s_f\right)\right]\,\right] du \qquad (5.1A)$$

Similarly,

$$E(e_E) = \frac{A}{A+1} m s_a \int_0^s \frac{1}{s}\,\frac{\lambda_E}{\lambda_E+\phi}\left[\,1-\exp\left[-(\lambda_E+\phi)\ \min\left(s-u,\ s_f\right)\right]\,\right] du \qquad (5.2A)$$

Under Case B: $s > s_a$:

$$E(e_c(s)|n_C) = n_C \int_0^{s_a} \frac{1}{s_a}\, P\left[Y_c < W_c,\, Y_c < s-u,\, Y_c < s_f\right] du$$

$$\therefore\ E(e_c) = \frac{1}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\,\frac{\lambda_C}{\lambda_C+\phi}\left[\,1-\exp\left[-(\lambda_C+\phi)\ \min\left(s-u,\ s_f\right)\right]\,\right] du \qquad (5.1B)$$

$$E(e_E) = \frac{A}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a}\,\frac{\lambda_E}{\lambda_E+\phi}\left[\,1-\exp\left[-(\lambda_E+\phi)\min\left(s-u,\ s_f\right)\right]\,\right] du \qquad (5.2B)$$

Under Case B', where real time $s = s_a + s_f$,

$$E(e_c)=\frac{1}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a} \frac{\lambda_C}{\lambda_C+\phi} \left[\, 1-\exp\left[-(\,\lambda_C+\phi)\min\left(s_a+s_f-u,\;\; s_f\right)\right]\,\right] du \qquad (5.1B')$$

$$E(e_E)=\frac{A}{A+1} m s_a \int_0^{s_a} \frac{1}{s_a} \frac{\lambda_E}{\lambda_E+\phi} \left[\, 1-\exp\left[-(\,\lambda_E+\phi)\;\min\left(s_a+s_f-u,\;\; s_f\right)\right]\,\right] du \qquad (5.2B')$$

The reason that we spent so much on deriving $\mathrm{E(e_c)}$ and $\mathrm{E(e_E)}$ is because: for a fixed sample design, to test $\mathrm{H_0\colon ln(\Delta) = 0\ vs.\, H_A\colon ln(\Delta) > 0}$, Appendix A1 of Rubinstein, Gail and Santer (1981) proved that $\ln(\widehat{\Delta})$ is asymptotically normally distributed with mean $\ln(\Delta)$ and variance $\sigma^2 = [\mathrm{E(e_c)}]^{-1} + [\mathrm{E(e_E)}]^{-1}$, where the total trial duration is $\mathrm{s_a +\ s_f}$. That is: $\sigma^2 = \mathrm{V(s_a + s_f)} = [\mathrm{E(e_c)}]^{-1} + [\mathrm{E(e_E)}]^{-1}$, where asymptotically being $4/\mathrm{d_{fix}}$, with $\mathrm{d_{fix}} = \mathrm{E(e_c)}$+$\mathrm{E(e_E)}$, the total number of events accumulated at time $\mathrm{s_a +\ s_f}$. Note that $\mathrm{E(e_E)},\ \mathrm{E(e_c)},\ \mathrm{V, d}$ are all function of time s on (0, $\mathrm{s_a +\ s_f}$], which can also be interchangeably represented as $\mathrm{E(e_E(s))}$, $\mathrm{E(e_c(s))},\ \mathrm{V(s)}$ and $\mathrm{d(s)}$.

Section 5.3.2: Survival Group Sequential Designs

For a group sequential design to test $H_0\colon \ln(\Delta) = 0\ \mathrm{vs.}\, H_A\colon \ln(\Delta) > 0$ with $i = 1,2, \ldots K,$ we have to satisfy both type I and II error requirements under a group sequential setting. Considering a group sequential trial with K planned analyses, let θ be the parameter of interest, a measure of placebo-drug difference and assume it can be estimated from trial data. The distribution of statistics Z_1 , Z_2 , …, Z_K are derived from cumulative data up to stages from $1{,}2\ \ldots, K,$ and it follows a canonical joint form (Chapter 3, Jennison and Turnbull, 2000) of multivariate normal distribution with E(Z_i) = $\theta\sqrt{t_i}$ and Cov(Z_i, Z_j) = $\sqrt{t_i/t_j}$, 1≤ $i \le j \le K$ and $\{t_1, \ldots, t_K\}$ are standard information levels for parameter θ, whith final $t_K = 1.$

Startng with notations in Section 5.2, where time s is on a continuous scale ranging from 0 to end of study time $s_a +\ s_f$, analysis times in group sequential design are discretized into K time points. Accordingly, to accommodate group sequential notations, we denote, on the discretized time points instead, $e_{c,i}$ as the accumulative number of events at Stage i, which is the same as $e_c(t_i)$ in Section 5.3.1. Similarly, $e_{E,i}, d_i, V_{i,}\mathrm{i} = 1, \dots, \mathrm{K},$ are discretized versions of $e_E(t_i), d(t_i)$ and $V(t_i)$, respectively, with $s = t_i$.

Because of asymptotic normality of standardized log-rank statistic (Chapter 13.2, Jennison and Turnbull), $\hat{\theta} = \frac{\ln(\hat{\Delta})}{\sqrt{\hat{\sigma}^2}}$ obtained at stage i aproximately has the canonical joint distribution. The standardized information level t_i also equals the ratio of variance accumulated at s_i relative to that of at the end of the trial ($s_a + s_f$). That is:

$$t_i = \frac{V_i}{V_K} = ([E(e_{c,i})]^{-1} + [E(e_{E,i})]^{-1})/([E(e_{c,K})]^{-1} + [E(e_{E,K})]^{-1}) \approx \frac{\frac{4}{d_i}}{\frac{4}{d_K}} \tag{5.3}$$

where observed information and required information (per group sequential theory) at time s_i are on the left and right sides of “approximately equal sign”(i.e., $' \approx '$), respectively.

For a group sequential test, upper efficacy boundaries $\{u_1, \dots, u_K\}$ (see Equation 5.4 below) are made to preserve type I error under null hypothesis. Non-binding boundaries $\{u_1, \dots, u_K\}$ are used in this paper as their calculations don’t depend on lower bounds $\{l_1, \dots, l_K\}$. Fisher’s information vector for a group sequential trial is searched to maintain per-specified power under alternative hypothesis (Equation 5.5); and in the end would equal to $R_{gsd} * \{t_1, \dots, t_K\}$ (Jennison and Turnbull, 2000).

$$P_{H_0}\{Z_1 \geq u_1 \cup Z_2 \geq u_2 \cup \cdots \cup Z_K \geq u_K\} = \frac{\alpha}{2} \tag{5.4}$$

$$P_{H_A}\{Z_1 \geq l_1\} + P_{H_A}\{l_1 \leq Z_1 \leq u_1, Z_2 \geq u_2\} + \cdots + P_{H_A}\{l_1 \leq Z_1 \leq u_1, \dots, l_{K-1} \leq Z_{K-1} \leq u_{K-1,} Z_K \geq u_K\} = 1 - \beta \tag{5.5}$$

Tables and Figures in this paper are created using Wang and Tsiatis (1987) (referred to as ‘WT’) with shape parameter of 0.15 for efficacy upper boundaries. Besides, for lower bounds $\{l_1, \ldots, l_K\}$, power spending is used with shape parameter of 0.8 (Kim and DeMets, 1987, referred to as 'Kim-DeMets'). That is: $f(t_i, \beta) = \beta * t_i^{0.8}, i = 1,2, \ldots, K$. For a equally spaced three-stage group sequential design (ie, $t = (0.33, 0.67,1)$), the cumulative type II error when overall $\beta = 0.2$ is $f(t, \beta) = (0.082, 0.145, 0.2)$.

Section 5.3.3: Operation Characteristics for Survival Group Sequential Trials with a Fixed Follow-up Period

Equation 5.6 below is the key equation to obtain real time of a survival group sequential trial with fixed follow-up time on every subject in the trial. To implement a particular group sequential test, Fisher's information for a group sequential trial is obtained by multiplying the Fisher's information of the fixed sample design by a factor to ensure power requirement (Jennison and Turnbull, 2002). Therefore, the variance of sequential test at time $\mathrm{t_i}$ is the time fraction multiplying $\mathrm{R_{gsd}}$, and then multiplying variance of the corresponding fixed sample design. Suppose analysis time s becomes $\mathrm{s_i}$, $= 1, \ldots, \mathrm{K}$, variance at $\mathrm{s_i}$ is:

$\mathrm{V(s) = \ t_i * R_{gsd} * \sigma_{fix}^2 = \frac{t_i * R_{gsd} * d_{fix}}{4}}$.

On the other hand, because variance of $\ln(\widehat{\Delta})$ at time s is

$\mathrm{V(s) = \ E\big(e_c(s)\big)]^{-1} + [E\big(e_E(s)\big)]^{-1}}$, resulting in information at real time s being $\mathrm{\frac{1}{E\big(e_c(s)\big)]^{-1} + [E\big(e_E(s)\big)]^{-1}}}$. Equating information collected in the trial at time of analysis and required information per group sequential theory, we have the following the key equation for calibrating operation characteristics for a survival trial with a fixed follow-up period:

$$\mathrm{\frac{1}{E\big(e_c(s)\big)]^{-1} + [E\big(e_E(s)\big)]^{-1}} = \frac{1}{t_{i*}\left(\frac{1}{4}\right) * d_{fix} * \ R_{gsd}}} \qquad (5.6)$$

Here are the steps to calculate design parameters for a group sequential trial for survival endpoints with a fixed follow-up period:

1) Use α, β and log hazard ratio under alternative hypothesis to calculate the required number of events d_{fix} for a fixed sample design.
2) With design parameters $\alpha, \beta,\ \{t_1, \dots, t_K\}$, upper efficacy boundaries (i.e., non-blinding WT with shape parameter of 0.15) together with Kim-DeMets (1987) lower boundaries with shape parameter of 0.8 , Equations 5.4 and 5.5 are utilized to calculate $\{l_1, \dots, l_K\}$, $\{u_1, \dots, u_K\}$, and R_{gsd}.
3) The required number of events at interim and final are then $d_{fix} * R_{gsd} * \{t_1, \dots, t_K\}$.
4) Given $\mathrm{s_f}$ (i.e., length of the fixed follow-up time), calculate needed accrual time $\mathrm{s_a}$ for a group sequential design to ensure power of group sequential test. This can be achieved by accumulating $\mathrm{d_{fix} * R_{gsd}}$ number of events at the end of the trial (i.e., at time of $\mathrm{s_a + s_f}$). That is: Set $\mathrm{t_i = 1}$ in Equation 5.6 and utilizes Equations 5.1B’ and 5.2B’ to obtain $\mathrm{E(e_c(s))}$ and $\mathrm{E(e_E(s))}$, respectively. Based on Equation 5.6 and making use of inverse-grid search, accrual time $\mathrm{s_a}$ for this group sequential trial is obtained.
5) For a range of accrual time $s \in [0.01, \mathrm{s_a}]$, with increment of 0.01 years, corresponding $\mathrm{E(e_c(s))}$ and $\mathrm{E(e_E(s))}$ can be calculated where Equations 5.1A and 5.2A are used when s$\leq \mathrm{s_a}$ and Equations 5.1B and 5.2B are used when $\mathrm{s > s_a}$. Real trial times, $\mathrm{s_i}$, for interim analysis are then obtained using inverse search to ensure information at interim analysis $\mathrm{i, i = 1, \dots, K-1}$ via Equation 6. Note that for the final analysis K, real time $\mathrm{s_K}$=$\mathrm{s_a + s_f}$ is already obtained in Step 4) above.
6) Number of patients to recruit at Stage $i, i = 1, \dots, K,$ is $N_i = ms_i$ if $s_i \leq s_a$, otherwise $N_i = ms_a$ if $s_i > s_a$.

In summary, the required maximum number of events is calculated based on group sequential theory to ensure enough power of detecting a hazard ratio of interest under alternative hypothesis while well-controlling of overall false positive rate. The accrual time for the whole group sequential trial s_a is calculated via obtaining enough information to achieve maximum information at the final analysis K. For interim analysis, at a real time after first-patient-in, events occurred up to it will be calculated via the pair of Equations 5.1A and 5.A, (or the pair of 5.1B and 5.2B, or the pair of 5.1B' and 5.2B') conditional upon the fact that event/censoring times are truncated above by s_f in the trial. And the real time for interim analysis can be reversely calculated by equating observed information so far with information needed at interim per group sequential asymptotic theory. Number of recruited patients at interim can thus be calculated with the help of accrual rate and real time at interim analysis (see Step 6 above).

Section 5.4: Examples

All examples use one-sided type I error of 0.025, power of 0.8, $K = 3$, and with median time of failure for the control group to be 1 year. Three different information times are chosen, as follows: $t^{(1)} = (0.33, 0.67, 1)$, $t^{(2)} = (0.5, 0.75, 1)$, and $t^{(3)} = (0.2, 0.8, 1)$ to represent equal increment of time fraction, interims occurring in the later part of the study, and first interim occurred in the early part and later ones in the later part, respectively.

Hazard ratio λ_c/λ_E is ranging from 1.3 to 3 in Figures 5.2 and 5.3. Lower rate of accrual with $m = 50$ per year is used to compare with brisk accrual of $m = 200$ per year (i.e., 17 patients per month). Three-stage group sequential WT designs together with fixed sample design(denoted as 'Fixed') are carefully investigated for the required accural time or total trial duration in the Tables 5.1 - 5.4 and Figures 5.2 - 5.3 regarding the following four categories:

Type A: with no censoring ($\phi = 0$) and short period of follow-up ($s_f = 0.5$ years)

Type B: With censoring($\phi = \lambda_c/2$) and short period of follow-up ($s_f = 0.5$ years)

Type C: with no censoring ($\phi = 0$) and long period of follow-up ($s_f = 1$ years)

Type D: With censoring($\phi = \lambda_c/2$) and long period of follow-up ($s_f = 1$ years)

In Figures 5.2 – 5.3, Types A, B, C and D are depicted using solid, medium dash, dash-dot and dotted line, respectively. Interestingly, they visually top each other in the order of B-A-D-C from upper- and right- most to lower- and left- most in the graphs. Comparing Type B with Type A, as well as Type D vs. Type C, shows the impact of competitive censoring on enlarging necessary accrual time and trial duration. The long length of follow-up period on shortening accrual time is shown via comparing designs having $s_f = 1$ years with those having $s_f = 0.5$ years. The impacts of varying hazard ratios and slow accrual versus quick enrolment rate on trial planning are assessed by evaluating Types A, B, C and D under a certain combination of hazard ratio and accrual rate.

Table 5.1 shows that eliminating censoring decreases required accrual time more for low accrual rate than for high accrual rate: under $t^{(1)}$**,** by 3.68 years for WT with rate of 50 per year and hazard ratio of 1.3 (from 47.21 years to 43.53 years), while only 0.92 years (from 11.79 years to 10.87 years) for rate of 200 per year at the same low hazard ratio of 1.3; similarly but in a much less extent for high hazard ratio of 3: by 0.44 years (from 5.51 years to 5.07 years) for m = 50 per year as compared with by 0.11 years (from 1.36 years to 1.25 years) for m = 200 per year. Similar trends exist in all group sequential trials with three time information vectors as well as in fixed sample design.

When accrual rate is low and hazard ratio is small, much longer time is needed to accumulate events to ensure power, with which sometimes is unreasonably long and seems not feasible as a real trial that could possibly be conducted by humankind. Fortunately, either reasonable increase in accrual rate or increase in hazard ratio can shorten it up. For example, accrual time for WT designs with $t^{(1)}$ information time, in the presence of censoring $\phi = 0.5\lambda_C$, and every subject will be followed for one year is 29.01 years for $m = 50$ per year and $\Delta = 1.3$; 3.09 years for $m = 50$ per year and $\Delta = 3$; 7.24 years for $m = 200$ per year and $\Delta = 1.3$ and only 0.77 (i.e., the shortest) years for $m = 200$ per year and $\Delta = 3.0$. Given operational feasibility of multi-national (regional) trials in current practice, accrual 200 patients world-wide in a year is achievable. And due to large span of required accrual times for different combinations of accrual time, hazard ratio and follow-up time from our exercises, feasibility explorations should be carefully done at the stage of designing a trial prior to recruiting first patient, rather than starting a trial with whatever accrual rate at hand and passively waiting for events to occur. In the later case, the study team might have to wait forever to collect the targeted number of events, which was actually happening in one of the bipolar trials the author has worked at.

Table 5.1 shows that including one year of follow-up has shortened the required accrual years as compared with short follow-up period of 0.5 years for all subjects: from 43.53 to 24.93 years, from 5.07 to 2.62 years, from 10.87 to 6.21 years and from 1.25 to 0.65 years for WT tests performed at $\mathrm{t}^{(1)}$ information times in the absence of censoring with m=50 per year and $\Delta = 1.3$, m=50 per year and $\Delta = 3.0$, m=200 per year and $\Delta = 1.3$ and m=200 per year and $\Delta = 3.0$, respectively, where the saving in the last case with both high accrual rate and high hazard ratio is 48%!! Similar observations are also noticed in corresponding cases when censoring is indeed present.

As for designs under different information vectors, WT designs with $t^{(3)}$ generally have the shortest accrual times as compared with those both under $t^{(1)}$ and $t^{(2)}$ because stopping at the first interim, which is only 0.2 of the total information time (i.e., $t^{(3)}$), shortens the overall accrual time. And all three information vectors tend to have accrual times in a magnitude close to each other when both accrual rate and hazard ratio are high (i.e., m=200 per year and $\Delta = 3.0$) because the required number of events can be accumulated quick enough, in rates almost non-differentiable. WT designs with $t^{(2)}$, accordingly to Table 5.1, always have the largest accrual period among all cases (Table 5.1).

In the past two decades, whenever group sequential trials are mentioned, it is said that they apply for trials with slow accrual. However, due to rapid change in information technology and improvement in trial conducts, data cleaning and analysis can be accurately executed within 4-6 weeks in pharmaceutical companies and thus expand the use of group sequential designs in drug development for trials with a quick accrual. Further, adding a fixed follow-up period for all subjects in group sequential survival trials will subsequently increase accrual time comparing with fixed sample designs, regardless of the accrual rate, which eases operational requirement in time a little.

Table 20(Tab. 5.1): Accrual time for group sequential designs

Table 5.1: Accrual time for group sequential designs under different combinations of hazard ratio (low 1.3 vs. high 3.0) and accrual rate (slow 50 per year vs. brisk 200 per year) when WT boundary is used for upper efficacy with shape parameter of 0.15 and lower boundary of Kim-Demets for futility with shape parameter of 0.8, $\boldsymbol{\alpha = 0.025}$ and $\boldsymbol{\beta = 0.2}$.

		Fixed				WT			
		$\phi = 0$ $s_f = 0.5$	$\phi = 0.5\lambda_C$ $s_f = 0.5$	$\phi = 0$ $s_f = 1$	$\phi = 0.5\lambda_C$ $s_f = 1$	$\phi = 0$ $s_f = 0.5$	$\phi = 0.5\lambda_C$ $s_f = 0.5$	$\phi = 0$ $s_f = 1$	$\phi = 0.5\lambda_C$ $s_f = 1$
a= 50 $\boldsymbol{\Delta = 1.3}$	$t^{(1)}$	35.06	38.02	20.15	23.44	43.53	47.21	24.93	29.01
	$t^{(2)}$					43.94	47.66	25.16	29.28
	$t^{(3)}$					43.36	47.03	24.83	28.90
a = 50 $\boldsymbol{\Delta = 3.0}$	$t^{(1)}$	3.27	3.55	1.77	2.08	5.07	5.51	2.62	3.09
	$t^{(2)}$					5.12	5.56	2.65	3.12
	$t^{(3)}$					5.05	5.49	2.61	3.08
a = 200 $\boldsymbol{\Delta = 1.3}$	$t^{(1)}$	8.76	9.51	5.03	5.86	10.87	11.79	6.21	7.24
	$t^{(2)}$					10.98	11.91	6.27	7.31
	$t^{(3)}$					10.83	11.75	6.19	7.21

a = 200 **Δ = 3.0**	$t^{(1)}$	0.81	0.88	0.44	0.52	1.25	1.36	0.65	0.77
	$t^{(2)}$					1.26	1.37	0.65	0.78
	$t^{(3)}$					1.25	1.35	0.64	0.77

In Figures 5.2 -5.3, accrual time s_a required to conduct a test against $H_0\colon \ln(\Delta) = 0$ is plotted on the x- axis with size $\alpha = 0.025$ and power of 0.8 ($\beta = 0.2$) to detect the alternative Δ on the y-axis. Median time to failure for control group subjects is always 1 year. Figure 5.2 plots the curves for long duration trials with slow accrual (m = 50 per year) while Figure 5.3 plots short duration with a brisk accrual (m = 200 per year). Within each set (one particular design with a certain information time vector), consisting with four types, the uppermost curve represents Type B, the case with moderate censoring present and short follow-up period ($\phi = \lambda_c/2$ and $s_f = 0.5$ years); the second upper curve represents Type A, the case with no censoring and short follow-up period ($\phi = 0$ and $s_f = 0.5$ years); the second to the lowest curve represents Type D, the case with moderate censoring and one-year follow-up period for all subjects ($\phi = \lambda_c/2$ and $s_f = 1$ years); and the lowermost curve represents Type C, the case with no censoring and 1-year follow-up ($\phi = 0$ and $s_f = 1$ years). For any hazard ratio, the required accrual length to detect treatment difference will have a order of Type C<Type D<Type A<Type B, showing the need of more accrual time resulted from censoring process while on the contrary shortening accrual period when the accrual rate increases. And the separation between the pair A and B and the pair C and D shows that the impact on the accrual time from accrual rate change is more dramatic as compared with that of introducing competitive censoring process. In Figures 5.2 – 5.3, the upper left, upper right, lower left and lower right graphs are for fixed sample design, WT under $t^{(1)}$, $t^{(2)}$ and $t^{(3)}$, respectively. Figures 5.2 – 5.3 are the complete version of Table 5.1 with regard to the varying hazard ratio, which in all scenarios show a decrease function of the

required length of accrual time of the trial in the increase of hazard ratio (i.e., from 1.3 to 3.0). A much longer accrual time is required when a small hazard ratio is in need to detect treatment difference, which further emphasizes how important it is to explore design characteristics prior to trial start as well as during the trial for necessary sample size re-estimation in the middle of a trial if the design parameter is over-estimated beforehand to avoid a underpowered study. Comparing Figure 5.3 with Figure 5.2, accrual time for both fixed sample design and group sequential design with brisk accrual is much shortened up; and the impact of adding competition from censoring on accrual time tends to diminish but not disappear in Figure 5.3 when having a much higher accrual rate of m = 200 per year.

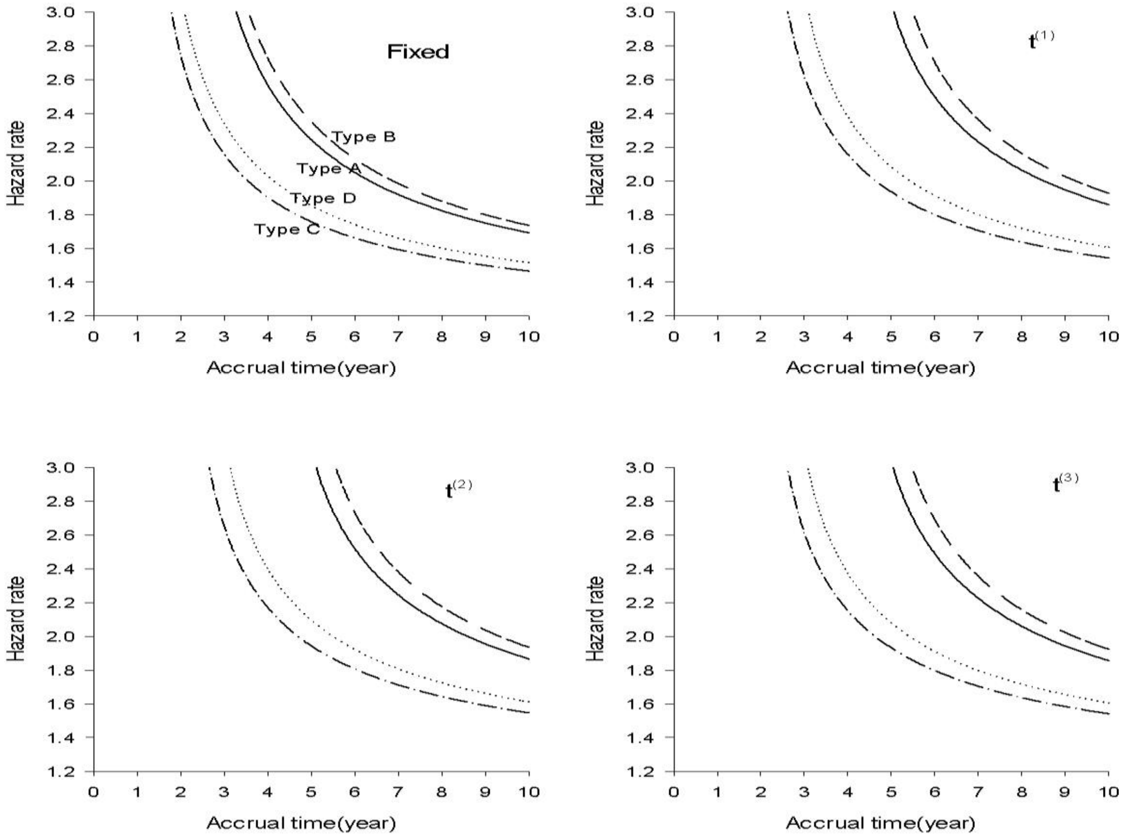


Figure 15(Fig. 5.2): Required accrual time (slow) vs. hazard ratio

Figure 5.2: Required accrual time vs. hazard ratio (from 1.3 to 3.0) for accrual rate of 50 per year, alpha=0.025, and beta=0.2

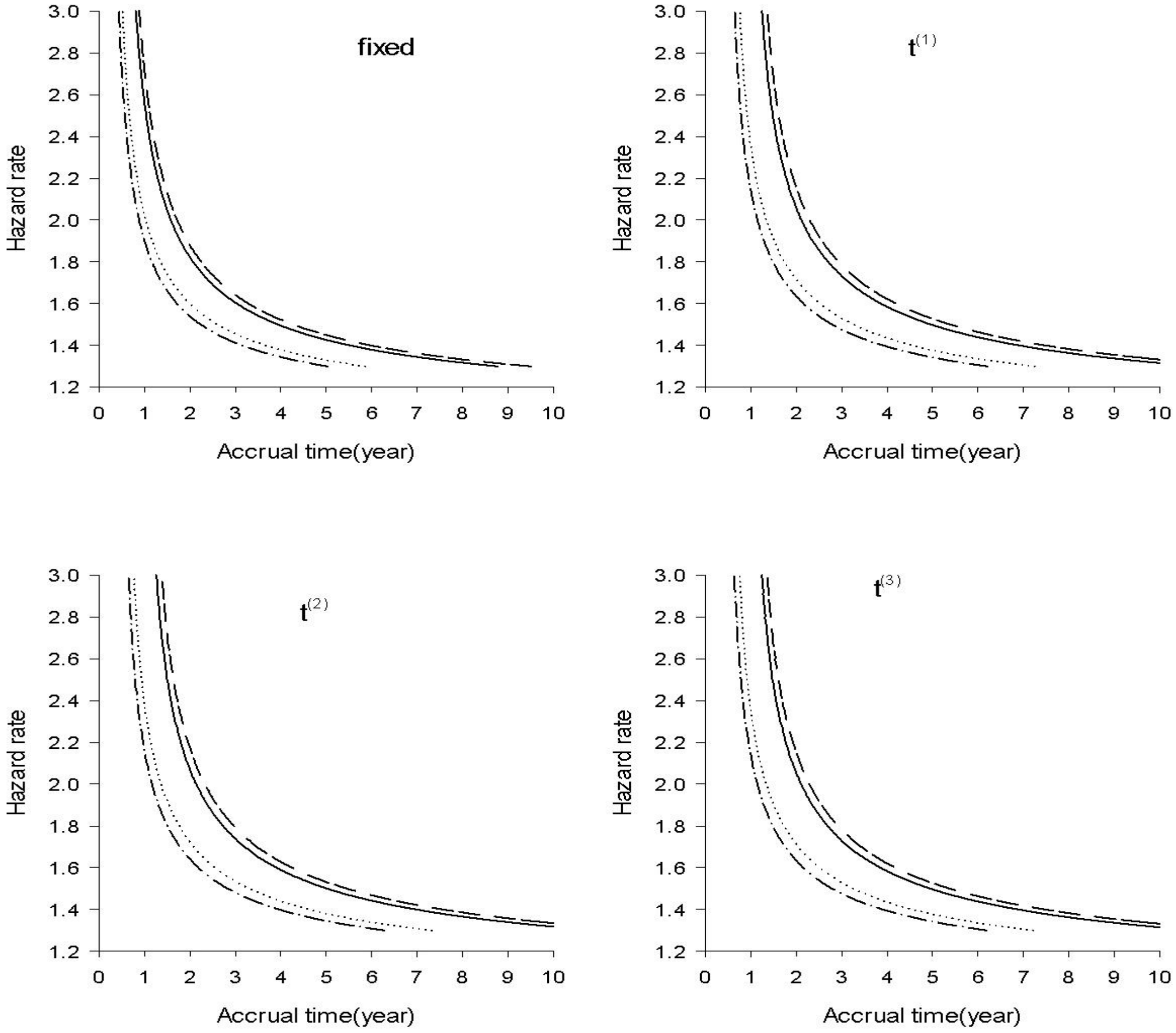


Figure 16(Fig. 5.3): Required accrual time (fast) vs. hazard ratio

Figure 5.3: Required accrual time vs. hazard ratio (from 1.3 to 3.0) for accrual rate of 200 per year, alpha=0.025, and beta=0.2

Besides accrual time length, total trial duration, which is the accrual time plus the follow-up time, is also investigated. In Tables 5.2 – 5.4, under $t^{(1)}$, $t^{(2)}$ and $t^{(3)}$ are, respectively, examined for four censoring rates of $\emptyset = 0$, $0.25\lambda_c$, $0.5\lambda_c$and λ_c, four follow-up times of $s_f =$ 0.5, 1, 1.5 and 2 years; and slow and brisk accrual rates of 50 per year and 200 per year as before, aiming at showing the magnitude of impact on total trial duration for a survival trial with different combinations of follow-up time, accrual rate and tested relative difference between placebo and treatment using WT upper boundary and Kim-DeMets lower boundary. For example, under $t^{(1)}$ and $\Delta = 2$ and s_f=1 year (shaded row in Table 5.2), a case embroils a real testing in drug development, the required total trial duration is 5.65, 6.03, 6.43 and 7.31 years for $\emptyset = 0$, $0.25\lambda_c$, $0.5\lambda_c$ and λ_c, respectively, with slow accrual of 50 patients per year while being 2.15, 2.25, 2.35 and 2.57 years correspondingly for fast accrual rate of 200 per year. There are indeed two ways to collect events quicker in a survival trial, recruiting more patients and following patients in the trial for a longer time. When comparing long follow-up time (i.e., s_f = 1 year) versus short follow-up time (i.e., s_f = 0.5 years), eliminating 0.5 years of follow-up length increases very little (i.e., 0.46 years) in total trial duration for a short duration trial with a rapid accrual, i.e., m = 200 per year, from 2.15 years to 2.61 years for $t^{(1)}$, $\Delta = 2.0$ and $\emptyset = 0$; but the recruited number of subjects will change from 178 patients (i.e., (2.15-1)*200 = 230) for s_f = 1 to 340 patients for s_f=0.5 (i.e., (2.61-0.5)*200 = 422). In other words 0.5 years shortening-up of follow-up time will result in accrual of an additional large chunk of patients (i.e., 92 more patients) and a longer trial (i.e., 0.46 years) to compensate for the shortened-up follow-up time 0.5 years.

Tables 5.2 – 5.4 furthermore show that, in contrast to long trial with slow accrual (m=50 per year), for short trials with rapid accrual rate (i.e., m = 200 per year), adding censoring process will increase accrual time, subsequently in total time to a less extent. Let's take $t^{(\mathbf{1})}$**,** $\Delta = 2.0, \mathrm{m} = 200$ per year, s_f = 1.0 years as an example, censoring ($\emptyset$ = $0.5\lambda_c$) adds 0.20 years in accrual time (from 2.15 years to 2.35 years) while for 0.78 years (from 5.56 years to 6.43 years) when with a shorter trial associated with low accrual time of m = 50 per year. Actually, from Figures 5.2 – 5.3, we can also see adding censoring changes little in accrual time for long trials with brick accrual unless hazard ratio is less than 2. On the other hand, this reminds us that accounting for censoring in design group sequential survival trials are important when we have a long trial associated with slow accrual and/or hazard ratio is small. In such cases, ignoring censoring will result in underestimated trial accrual time and total trial duration, which leads to inadequate design preparations.

Table 21(Tab. 5.2): Total trial duration for WT (shape = 0.15) group sequential trials

Table 5.2: Total trial duration for WT (shape = 0.15) group sequential trials when information vector is $t^{(1)}$, plus alpha=0.025, and beta = 0.2.

	Δ	∅=0		∅=0.25λ_c		∅=0.5λ_c		∅=λ_c	
		50	200	50	200	50	200	50	200
s_f=0.5	1.25	50.49	15.20	50.49	15.81	50.50	16.44	50.50	17.74
	1.5	20.44	5.47	21.28	5.68	22.13	5.90	23.91	6.35
	2	9.01	2.61	9.37	2.71	9.74	2.80	10.51	3.00
	3	5.57	1.76	5.79	1.81	6.01	1.86	6.48	1.99
s_f=1.0	1.25	34.82	9.45	37.53	10.13	40.34	10.82	46.33	12.33
	1.5	12.25	3.80	13.16	4.04	14.11	4.26	16.14	4.78
	2	5.65	2.15	6.03	2.25	6.43	2.35	7.31	2.57
	3	3.62	1.64	3.86	1.70	4.10	1.77	4.61	1.90
s_f=1.5	1.25	27.25	7.93	30.19	8.67	33.29	9.43	39.94	11.10
	1.5	9.93	3.60	10.92	3.85	11.95	4.11	14.21	4.67
	2	4.88	2.33	5.29	2.44	5.74	2.55	6.69	2.79
	3	3.33	1.94	3.56	2.00	3.82	2.07	4.38	2.21
s_f=2.0	1.25	23.90	7.47	27.05	8.26	30.40	9.09	37.61	10.89
	1.5	9.07	3.76	10.13	4.03	11.25	4.30	13.69	4.91
	2	4.77	2.68	5.20	2.80	5.68	2.91	6.70	3.16
	3	3.44	2.34	3.69	2.41	3.95	2.47	4.55	2.63

Table 22(Tab. 5.3): Total trial duration for WT (shape = 0.15) group sequential trials

Table 5.3: Total trial duration for WT (shape = 0.15) group sequential trials when information vector is $t^{(2)}$, plus alpha = 0.025, and beta = 0.2.

	Δ	∅=0		∅=0.25λ_c		∅=0.5λ_c		∅=λ_c	
		50	200	50	200	50	200	50	200
s_f=0.5	1.25	50.49	15.34	50.49	15.95	50.50	16.59	50.50	17.91
	1.5	20.63	5.52	21.47	5.73	22.34	5.95	24.13	6.40
	2	9.09	2.63	9.45	2.73	9.83	2.82	10.60	3.02
	3	5.62	1.77	5.84	1.82	6.06	1.87	6.53	2.01
s_f=1.0	1.25	35.14	9.53	37.88	10.22	40.71	10.92	46.76	12.43
	1.5	12.36	3.83	13.28	4.06	14.24	4.29	16.28	4.81
	2	5.696	2.16	6.08	2.26	6.49	2.36	7.37	2.58
	3	3.652	1.65	3.89	1.71	4.13	1.78	4.65	1.90
s_f=1.	1.25	27.50	7.99	30.46	8.74	33.59	9.51	40.30	11.19

5	1.5	10.01	3.62	11.01	3.87	12.05	4.14	14.33	4.70
	2	4.91	2.34	5.32	2.45	5.78	2.56	6.74	2.80
	3	3.35	1.95	3.58	2.01	3.85	2.08	4.40	2.22
s_f=2.0	1.25	24.11	7.52	27.29	8.32	30.67	9.16	37.95	10.98
	1.5	9.14	3.78	10.21	4.05	11.34	4.32	13.780	4.94
	2	4.80	2.69	5.23	2.80	5.71	2.92	6.74	3.17
	3	3.45	2.35	3.701	2.41	3.97	2.48	4.58	2.63

Table 23(Tab. 5.4): Total trial duration for WT (shape = 0.15) group sequential trials

Table 5.4: **Total trial duration for WT (shape = 0.15) group sequential trials when information vector is $t^{(3)}$, plus alpha = 0.025, and beta = 0.2.**

	Δ	∅=0		∅=0.25λ_c		∅=0.5λ_c		∅=λ_c	
		50	200	50	200	50	200	50	200
s_f=0.5	1.25	50.49	15.15	50.49	15.75	50.50	16.38	50.50	17.68
	1.5	20.36	5.45	21.20	5.66	22.05	5.88	23.82	6.32
	2	8.98	2.60	9.34	2.70	9.70	2.79	10.47	2.99
	3	5.55	1.75	5.77	1.80	5.99	1.85	6.45	1.99
s_f=1.0	1.25	34.70	9.42	37.39	10.09	40.19	10.79	46.16	12.28
	1.5	12.21	3.79	13.12	4.02	14.06	4.25	16.08	4.76
	2	5.63	2.14	6.02	2.24	6.41	2.34	7.28	2.56
	3	3.61	1.64	3.85	1.70	4.09	1.77	4.60	1.89
s_f=1.5	1.25	27.16	7.91	30.08	8.64	33.17	9.40	39.79	11.06
	1.5	9.89	3.59	10.88	3.84	11.92	4.11	14.16	4.66
	2	4.86	2.33	5.27	2.43	5.72	2.55	6.67	2.78
	3	3.32	1.94	3.55	2.00	3.81	2.07	4.37	2.21
s_f=2.0	1.25	23.82	7.45	26.96	8.23	30.29	9.06	37.48	10.86
	1.5	9.04	3.76	10.10	4.02	11.21	4.29	13.64	4.90
	2	4.76	2.68	5.19	2.79	5.66	2.91	6.68	3.16
	3	3.43	2.34	3.68	2.41	3.95	2.47	4.55	2.62

Based on required number of events for a group sequential design, accrual time and total trial duration for survival group sequential trial with fixed follow-up time can be derived. Impacts from censoring and different follow-up periods are addressed above in Tables 5.1 – 5.4 and Figures 5.2 – 5.3. There are three other aspects of group sequential designs that needed to be explored prior to trial start, as interim analyses allowing for early stopping using results from accumulating data up to analysis stage in contrast to fixed duration fixed sample design. These parameters are as follows:

i) Real time at interim and final analyses;

ii) Required number of events at each analysis including interim and final;

iii) Accrued number of patients at each analysis including interim and final.

As described above, inverse searching utilizing numerical integration is implemented to find the real time for each analysis; then accumulated number of patients at time is calculated to accumulate required number of events at each analysis so that overall power of detecting treatment effect is ensured. One moderate hazard ratio, i.e., $\Delta = 2$, is picked up to tabulate the operation characteristics group sequential trials with WT upper boundary and Kim-DeMets lower boundary. Tables 5.5 – 5.6 list design specifics which re-emphasize the impact of censoring and length of follow-up period on trial designs. Besides new features like number of patients and real time at interim, other group sequential parameters like upper and lower bounds are also tabulated. Probability and expected information under null or alternative are not included due to space limitation.

Tables 5.5 and 5.6 depict operation characteristics for designs with follow-up time of 0.5 or 1 years, and under $t^{(1)}$, $t^{(2)}$ or $t^{(3)}$. In each table, there are four cases in combination of censoring status and an accrual rate (50 per year or 200 per year):

Case I: $\phi = 0$ and m = 50 per year;

Case II: $\phi = 0.5\lambda_c$ and m = 50 per year;

Case III: $\phi = 0$ and m = 200 per year;

Case IV: $\phi = 0.5\lambda_c$ and m = 200 per year.

Using asymmetric three-stage group sequential design, under equally-spaced $t^{(1)}$, the upper WT boundaries with shape parameter of 0.15 is $u = (3.009054, 2.348463, 2.041314)$ and Kim-Demets lower boundaries with shape parameter of 0.8 is $l = (0.392837, 1.288984, 2.041314)$.

The trial will stop for efficacy if log-rank test statistic is greater than or equal to 3.009054 at first stage or greater than or equal to 2.348463 at the second stage, stop for futility if less than 0.392837 at Stage One or less than 1.288984 at Stage Two; and at the final stage will reject null if logrank test statistic is greater than or equal to 2.041314 and accept otherwise. The required number of events to conduct analysis is 28, 56 and 88 at Stage One, Stage Two and the final stage, respectively. From Table 5, for fixed follow-up of 0.5 years for each subject and in the absence of censoring, the first interim analysis will occur at 3 years after date of first-patient-in (denoted as 'FPI') with 150 patients accrued in the trial for accrual rate of 50 per year (i.e., Case I under $t^{(1)}$ in Table 5.5) while around 0.9 years after FPI with 180 patients accumulated for accrual rate of 200 per year (i.e., Case III under $t^{(1)}$ in Table 5.5); the second interim will occur at 5.90 years with 295 patients accumulated in the trial and 1.65 years with 330 patients accrued for accrual rate of 50 per year and 200 per year, respectively. Subsequently, the final analysis will occur at 9.01 years with 425 patients accrued in total and 2.63 years with the same amount of subjects accumulated, under which the accrual time for slow and fast accruals respectively has to recruit subjects for 8.51 years and 2.13 years. In the presence of censoring, accordingly Case II and IV in Table 5.5, accrual time, subsequently total trial duration and recruited number of patients will all increase in order to accumulate the same number of events comparing trial that in the absence of censoring for detecting the same alternative hypothesis of $\Delta = 2$.

Comparing operation characteristics for short follow-up time with long follow-up time (Table 5.5 vs. Table 5.6), under $\mathrm{t}^{(1)}$, in Case I of slow accrual in the absence of censoring, adding 0.5 years of follow-up for each subjects resulted in saving of 3.86 years (45%) in accrual time (from 8.51 years to 4.65 years), saving of 3.46 (38%) in total trial length (from 9.01 years to 5.65

years) and saving of 193 (45%) in accrued number of patients (from 425 to 232) to test against equality of hazard rate when trial is powered at hazard ratio of 2. Additionally, for fast accrual and long follow-up trials, i.e., Case III and IV in Table 5.6, there is no need to recruit patients after Interim Two as enough patients have been recruited at time of Interim Two; and the trial team can stop enrollment and wait patiently for more events to occur for the final stage and then terminate the trial. Therefore, without exploration of trial operation characteristics, the study team has no way be aware of when to stop enrollment of patients and when to get preparations done upon the right timing for interim and final analyses in group sequential survival trials with fixed length of follow-up time; and neither do they know how to adjust these parameters when accrual rate changes during the trial and the extent of censoring is different from what they thought prior to trial start.

Table 24(Tab. 5.5): Group sequential designs

Table 5.5: Group sequential design with WT upper bounds (shape=0.15) and Kim-Demets beta-spending lower bounds with shape parameter of 0.8, alpha=0.025, beta=0.2, hazard ratio=2 and $s_f = 0.5$ while Case I: $\phi = 0$ and m=50 per year; Case II: $\phi = 0.5\lambda_c$ and m=50 per year; Case III: $\phi = 0$ and m=200 per year; and Case IV: $\phi = 0.5\lambda_c$ and m=200 per year.

	# of events	Information time	bounds		Real time (year)				Number of Patients				Accrual time / follow-up time (year)			
			l	u	Case I	Case II	Case III	Case IV	Case I	Case II	Case III	Case IV	Case I	Case II	Case III	Case IV
$t^{(1)}$	28	0.33	0.392837	3.009054	3.00	3.25	0.9	0.95	150	162	180	190	8.51/0.5	9.24/0.5	2.13/0.5	2.30/0.5
	56	0.67	1.288984	2.348463	5.90	6.40	1.65	1.75	295	320	330	350				
	88	1.0	2.041314	2.041314	9.01	9.74	2.63	2.80	425	462	425	461				
$t^{(2)}$	44	0.5	1.002881	2.631239	4.50	4.85	1.30	1.35	225	243	260	270	8.59/0.5	9.33/0.5	2.15/0.5	2.32/0.5
	66	0.75	1.479783	2.283118	6.65	7.20	1.80	1.95	333	360	360	390				
	89	1.0	2.064428	2.064428	9.09	9.83	2.65	2.82	429	466	429	465				
$t^{(3)}$	17	0.2	-0.21179	3.59089	1.90	2.05	0.65	0.65	95	103	130	130	8.48/0.5	9.20/0.5	2.12/0.5	2.29/0.5
	70	0.8	1.678852	2.210452	7.00	7.60	1.90	2..05	350	380	380	410				
	87	1.0	2.044384	2.044384	8.98	9.70	2.62	2.79	424	460	423	459				

Table 25(Tab. 5.6): Group sequential designs

Table 5.6: Group sequential design with WT upper bounds (shape=0.15) and Kim-Demets beta-spending lower bounds with shape parameter of 0.8, alpha=0.025, beta=0.2, hazard ratio=2 and $s_f = 1.0$ while Case I: $\phi = 0$ and m=50 per year; Case II: $\phi = 0.5\lambda_c$ and m=50 per year; Case III: $\phi = 0$ and m=200 per year; and Case IV: $\phi = 0.5\lambda_c$ and m=200 per year.

	# of events	Information time	bounds		Real time (year)				Number of Patients				Accrual time / follow-up time (year)			
			l	u	Case I	Case II	Case III	Case IV	Case I	Case II	Case III	Case IV	Case I	Case II	Case III	Case IV
$t^{(1)}$	28	0.33	0.392837	3.009054	1.95	2.20	0.80	0.85	98	110	160	170	4.65/1.0	5.44/1.0	1.16/1.0	1.35/1.0
	58	0.67	1.288984	2.348463	3.55	4.05	1.20	1.30	178	203	231	260				
	86	1.0	2.041314	2.041314	5.65	6.44	2.16	2.35	232	272	231	270				
$t^{(2)}$	43	0.5	1.002881	2.631239	2.80	3.15	1.00	1.10	140	158	200	220	5.69/1.0	5.49/1.0	1.17/1.0	1.36/1.0
	65	0.75	1.479783	2.283118	3.95	4.50	1.35	1.45	198	225	233	272				
	87	1.0	2.064428	2.064428	5.69	6.49	2.17	2.36	234	274	233	272				
$t^{(3)}$	17	0.2	-0.21179	3.59089	1.35	1.50	0.60	0.65	68	75	120	130	4.63/1.0	5.41/1.0	1.16/1.0	1.34/1.0
	68	0.8	1.678852	2.210452	4.15	4.75	1.40	1.50	208	238	231	268				
	86	1.0	2.044384	2.044384	5.63	6.41	2.16	2.34	231	270	231	268				

Section 5.5: Discussion

Randomized clinical trials have been widely used in clinical trial submissions to assess maintenance effect of investigational compound relative to placebo in the double-blind phase on patients who have been stabilized for symptoms after a period of open-label treatment phase. For a trial design without a fixed follow-up period for each subject as in Figure 5.1a, randomized subjects are followed-up until event occurring, or early withdrawal, or until trial termination, whichever date comes the earliest. There are issues observed from drug development practice in trials without a fixed follow-up length imposed on all subjects as follows: safety parameters can't be interpreted properly due to variable duration in the double-blind phase; some overtime effects measured by scales using repeated measures can't be evaluated properly because missing is not at random; and long exposure to the investigation medication of those patients remaining until study termination is also questionable. Adding a fixed-length of follow-up time for all subjects can alleviate above issues in certain extent as discussed in our motivation example (Section 5.1). Especially, for trials comparing investigational drug against active comparator when relapse rates are low in both groups so that most of subjects in the double-blind phase will be administratively censored at the end of the follow-up time with time to censoring s_f, safety and secondary efficacy endpoints in this case can, in some extent, be assessed properly by having the same trial length among these subjects. In the meantime, primary efficacy endpoint can be addressed in a better way as compared with a trial without a fixed follow-up period, because in the intent-to-treat analysis, there can't exist a large chunk of subjects being administratively censored at the study termination with a minimum exposure up to one day to the study medication so that resulting in no attribution to evaluation of the overall treatment effect between two survival curves.

Careful explorations of accrual time requirement are needed prior to trial start, Table 5.1 shows that some trials are desperately long with slow accrual rate when to test small hazard ratio, which is often occurred in non-inferiority randomized withdrawal trials, as our motivation example, statistical exploration of trial feasibility is a must to predict large enough accrual rate to finish trial earlier especially in face of nowadays' fierce competition in drug development. Impacts of censoring can also be explored a priori as non-administrative censoring is determined to exist in every trial but in a different extent, which, by Tables 5.1 - 5.6 and Figures 5.2 – 5.3, is a factor to determine trial length and required number of patients. By our explorations, the length of follow-up time has substantial impacts on trial accrual time as well on total trial duration and recruited patients' number. The minimum exposure length is normally chosen to account for the requirement of both safety and tolerability of study drug in balance with the need of long enough exposure to detect placebo-treat difference in efficacy. Within a range of fixed follow-up lengths, which are all longer than the minimum exposure requirement and under which subjects are well-tolerated, a longer follow-up length can substantially save time and budget and can gather a better safety profile as compared with that of a short follow-up time. Additionally, real time for interim gives trial team in good preparation in time and is operationally highly appreciated because this prediction can avoid allocate resources too early or too late. Lastly, Newton-Rapshon search as used in Kim and Tsiatis (1990) is not working here, as we have a minimum function in the integrand part of the integration. Brutal force grid-search is proposed in the trial, but can be done very quickly even with a personal laptop. Of note, although our motivating example is a double-blind randomized withdrawal trial, methods established in this paper apply to any survival group sequential trials with a fixed follow-up period imposed on all subjects irrespectively of blind or open-label, maintenance study or direct confirmative study on drug efficacy in acute patients. It is also of note that subjects can still withdraw early from the trial prior to the maximum follow-up time if it is deemed necessary, because as pointed out by the reviewer that it may be equally unethical to force subjects to be studied by the same length if a subject changes the informed consent or encounters an unexpected adverse event.

Although Software ADDPLAN® and Software EAST® has implemented group sequential design for survival data and SAS® has SEQDESIGN and SEQTEST procedures to deal with designs and analyses, there hasn't been any publication substantively assessing the impacts of imposing a maximum follow-up period for each subject on operation characteristics of a particular design. This paper serves this purpose and furthermore, optimality feature could be assessed using automated written codes but hard to achieve using available software. Programming codes were done in R and available for distribution from the author upon request.

References

Armitage P. *Sequential Medical Trials*. Oxford: Blackwell. 1975.

Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B* 1972; 34(2): 187–220.

Gail MH, DeMets D L, Slud EV. Simulation studies on increments of the two sample logrank score test for survival time data, with application to group sequential boundaries. In *Survival analysis.* IMS Lecture Notes, Monograph Series 2, R. Johson and J. Crowley (eds), 287-301. Hayward, California: Institute of mathematical Statistics. 1981.

George SL, Desu MM. planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Disease* 1974; **27**:15-29.

Harrington D. Linear Rank Tests in Survival Analysis. *Encyclopedia of Biostatistics*. Wiley Interscience. 2005.

Jennison C , Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman and Hall. 2000.

Jennison C, Turnbull BW. Repeated confidence intervals for group sequential trials. *Controlled Clinical Trials* 1984; 5:33-45.

Kim, K. and Demets, D.L. (1987). Design and Analysis of group sequential tests based on Type I error spending rate functions. *Biometrika* 74, 149-54

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; **50**(3): 163–70.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35:549-56.

Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society, Series A* 1971; **135**(2): 185–207.

Rubinstein LV, Gail MH, Santer TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Disease* 1981; 34:469-479.

Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1977; 38:153-162.

Wang SK, Tsiatis AA, Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987, 43:193-199.

Sellke K, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983; 70:315-326.

Slud EV. Sequential linear rank tests for two-sample censored survival data. *Annuals of Statistics* 1984; **12**:551-571.

Slud EV, Wei LJ. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* 1982, **77**:862-868.

Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* 1982; **77**:855-861.

Chapter 6

Optimal Weighted *Z* Test and Linear Combination Test in Extended Sequential Parallel Designs

(Being reviewed by *Communication in Statistics: theory and method*)

Abstract: Many times in clinical trials using Sequential Parallel Design (SPD) with two treatments subjects are randomized in Period 1 and placebo non-responders are re-randomized in period 2 to either continue with placebo or switch to drug. In this paper, we introduce extended SPD (ESPD) and consider the re-randomization of not only placebo non-responders during Period 1 but also the re-randomization of drug responders during Period 1 into Period 2. Statistical methods to analyze data from an ESPD have been discussed. An optimal weighted Z test for normal data and a linear combination test for binary data are proposed and investigated.

Keywords: Weighted Z Test; Parallel Sequential Design; Double Randomization; Placebo Effect; Linear Combination Test**.**

Section 6.1: Introduction

To maintain the balance among baseline factors between treatment groups, randomization of subjects in different treatment groups is commonly used in randomized trials. After meeting inclusion/exclusion criteria, subjects are randomized onto either drug or placebo to assess drug effect. Given that baseline factors have been evenly balanced between comparing groups, observed drug-placebo difference can then be considered as a measure of drug effect on patient population. Although majority of clinical trials only have one randomization, there are occasions when subjects enter from first period to second period depending upon some success criteria and re-randomization is needed prior to subjects enter the Period 2. For instance, to investigate maintenance effect after having been stabilized on drug, the second randomization could eliminate the bias resulted from differential early withdrawals between groups. There is a rich history of published trials employing the double randomization in different therapeutic areas (Mills et al. 2007; Heyn et al. 1974; Habermann et al. 2006).

Strong placebo response has been problematic in central nervous system (CNS) clinical trials, leading to a reduced drug effect and thus resulting in decrease in probability of finding an effective drug (Khin et al. 2011). The ideal situation is to have comparative data collected only from subjects who are placebo non-responders. Stringent trial procedures together with enrichment of placebo non-responders are some of the ways to decrease placebo response in clinical trials. Fava et al. (2003) proposed a SPD where subjects are only randomized during Period 1. Accordingly, some placebo non-responders in Period 1 continue on placebo in Period 2 and others switch to drug in Period 2; and subjects who are treated with drug in Period 1 would continue to receive drug in Period 2. Treatment sequences for all subjects are all pre-specified prior to trial start; and data from Period 2 for subjects who are on drug in both periods are for safety evaluations only. An estimator is proposed to assess drug effect in each period, and a combined estimator is also proposed to test superiority of investigational drug over placebo across periods. Tamura & Huang (2007) suggest seemingly unrelated regression analysis (SUR) to obtain individual estimator from each period to analyze data from a SPD trial. To adjust for the bias caused by possible unbalanced dropouts among placebo non-responders in Period 1, both Fava et al. (2003) and Chen et al. (2011) propose re-randomizing Period 1 placebo non-responders into Period 2. They showed that when certain conditions are met the covariance between two estimators to be zero. Re-randomization of Period 1 placebo non-responders into Period 2 is also suggested by Liu et al. (2012) where they suggested a weighted Z test to increase efficiency of hypothesis test. This paper in addition to re-randomization of placebo non-responders in Period 1 also considers re-randomization of Period 1 drug responders into Period 2 after washing off the residual effects. Section 6.2 describes the design schematic, Section 6.3 introduces an optimal weighted Z test for normal data in an extended SPD trial and Section 6.4 proposes a linear combination test for binary data. Discussions and further research directions are provided in Section 6.5.

Section 6.2: Design Schematic

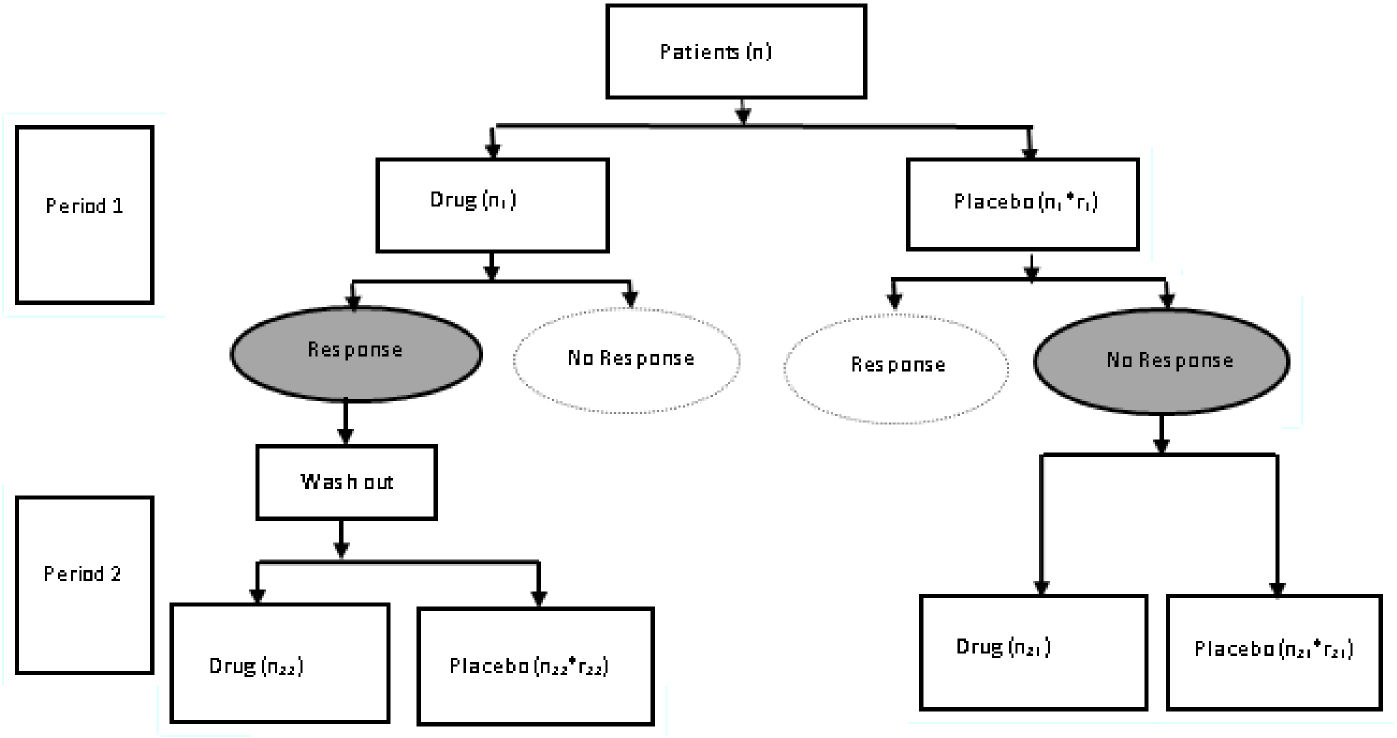


Figure 17(Fig. 6.1): Design schematic

Figure 6.1: Design schematic

Subjects with endpoint value of a period greater than or equal to a threshold value are defined as a responders during the period (or on the contrary, being less than or equal to a threshold value). The design consists of two periods. At the beginning of Period 1, eligible subjects are randomized to receive either placebo or drug, and subjects can withdraw early for lack-of-efficacy, adverse event, or other safety issues. At the end of Period 1, placebo patients are classified as responder or non-responder based on endpoint value. Placebo non-responders are re-randomized to receive either drug or placebo in Period 2. Similarly, subjects in drug group are also classified as responders or non-responders. A proper washout period is used to eliminate residue effects obtained from Period 1 and then drug responders are re-randomized to receive either placebo or continue on drug in Period 2. To maintain balance of baseline factors between comparing groups in Period 2, randomization ratio in Period 2 is set as 1:1 for both placebo non-responders and drug responders. Period 1 randomization ratio of 1:1 is not required but it should be 1:1 in Period 2 within each randomization group.

Section 6.3: Normal Data

Section 6.3.1: General Theory of Design

Let θ_1 be Period 1 drug effect with standard error υ_1. Pairs θ_{21} and υ_{21} are similarly defined for drug effect in Period 2 obtained from re-randomized Period 1 placebo non-responders, so do θ_{22} and υ_{22} obtained from re-randomized Period 1 drug responders. Let r_1 denote the randomization ratio of subjects receiving placebo versus drug in Period 1. Let r_{21} and r_{22} denote re-randomization ratio for placebo versus drug in Period 2 for Period 1 placebo non-responders and for Period 1 drug responders, respectively. Therefore, the number of subjects for drug and placebo in Period 1 are respectively n_1 and $n_1 * r_1$. Note that the sample sizes for both Period 1 placebo non-responders and drug responders are random and depend on the attrition rate in Period 1 as well as the probability of being a responder at the end of Period 1. n_{21}^{*} and $\mathrm{n}_{21}^{*} * \mathrm{r}_{21}$ are the expected number of Period 1 placebo non-responders who switch to receive drug in Period 2 and remain on placebo in Period 2, respectively. Similarly, n_{22}^{*} and $\mathrm{n}_{22}^{*} * \mathrm{r}_{22}$ are defined as the expected number of drug responders who remain on drug in Period 2 and switch to receive placebo in Period 2, respectively.

We are interested in testing the following global null hypothesis:

$H_0: \theta_1 \leq 0$ and $\theta_{21} \leq 0$ and $\theta_{22} \leq 0$ in favor of the alternative hypothesis

$H_A: \theta_1 > 0$ or $\theta_{21} > 0$ or $\theta_{22} > 0$

For θ_1, θ_{21} and θ_{22}, the test statistics for testing-the individual null hypothesis $H_{01}: \theta_1 \leq 0$, $H_{021}: \theta_{21} \leq 0$ and $H_{022}: \theta_{22} \leq 0$ are Z_1, Z_{21} and Z_{22}, respectively, with each test statistic defined as an estimate divided by its standard error. They are standard normal variables with mean zero and variance of one under null hypotheses and with a positive mean and variance of one under alternative hypothesis. Note that the individual statistics here are different from widely cited weighted Z statistic from two stages (Cui et al. 1999) resulting from a design with one randomization only. Here Z_{21} and Z_{22} are obtained from Period 2 after re-randomization. The relationships among Z_1, Z_{21} and Z_{22} are essential to understand asymptotical distribution of the combined test statistic under both null and alternative hypotheses. Since Period 1 placebo non-responders contribute to both Z_1 and Z_{21} and Period 1 drug responders contribute to both Z_1 and Z_{22}, correlation coefficient between them (i.e., Z_1 versus Z_{21} or Z_{22}) must be evaluated in order to test the hypothesis when using combined test statistic against the global null hypothesis. Let ρ_1 denote the correlation coefficient between outcomes at Period 1 and Period 2 for subjects who are placebo non-responders in Period 1 and then treated with placebo in Period 2 and ρ_2 is defined similarly but for subjects who are placebo non-responders in Period 1 and treated with drug in Period 2. Assuming equal correlation coefficients (i.e., $\rho_1 = \rho_2$) as in Chen et al. (2011), it is proved that covariance between $Z_1 \; and \; Z_{21}$ is zero (that is cov (Z_1, Z_{21}) = 0). Similarly, cov ($Z_1, \; Z_{22}$) = 0. Also Z_{21} and Z_{22} are independent of each other as coming from different cohorts of subjects in Period 2, which says cov ($Z_{21}, \; Z_{22}$)=0.

To establish the efficacy of the drug, one combines $Z_1, \; Z_{21}$and Z_{22} via

Z=$\sqrt{\lambda_1} Z_1 + \sqrt{\lambda_2} Z_{21} + \sqrt{1 - \lambda_1 - \lambda_2} Z_{22}$

Due to mutual independence, Var(Z) = 1 under both null or alternative hypotheses.—

with $R_k = \frac{r_k}{1+r_k}, \delta_k = \frac{\theta_k}{\upsilon_k}, for \; k = 1, 21 \; or \; 22$, one obtains

E(Z)= $\sqrt{\lambda_1}\sqrt{n_1 R_1}\delta_1 + \sqrt{\lambda_2}\sqrt{n_{21}^* R_{21}}\delta_{21} + \sqrt{1 - \lambda_1 - \lambda_2}\sqrt{n_{22}^* R_{22}}\delta_{22}$

This expectation is zero under null because having zero δ's under null. Furthermore, assuming positive δ's, maximizing the power of the test $Z > z_{1-\frac{\alpha}{2}}$ is equivalent to maximize the expectation of Z under alternative. Taking derivative of expectation function with respect to λ_1 and λ_2 separately, setting derivative function equal to zero and solving equations simultaneously, one can get optimal weights λ_1 and λ_2 as:

$$\lambda_1^* = \frac{n_1 R_1 \delta_1^2 n_{22}^* R_{22} \delta_{22}^2}{(n_1 R_1 \delta_1^2 + n_{22}^* R_{22} \delta_{22}^2)(n_{21}^* R_{21} \delta_{21}^2 + n_{22}^* R_{22} \delta_{22}^2) - n_1 R_1 \delta_1^2 n_{21}^* R_{21} \delta_{21}^2}$$

$$\lambda_2^* = \frac{n_{21}^* R_{21} \delta_{21}^2 n_{22}^* R_{22} \delta_{22}^2}{(n_1 R_1 \delta_1^2 + n_{22}^* R_{22} \delta_{22}^2)(n_{21}^* R_{21} \delta_{21}^2 + n_{22}^* R_{22} \delta_{22}^2) - n_1 R_1 \delta_1^2 n_{21}^* R_{21} \delta_{21}^2}$$

$$1 - \lambda_1^* - \lambda_2^* = \frac{n_{22}^* R_{22} \delta_{22}^2 n_{22}^* R_{22} \delta_{22}^2}{(n_1 R_1 \delta_1^2 + n_{22}^* R_{22} \delta_{22}^2)(n_{21}^* R_{21} \delta_{21}^2 + n_{22}^* R_{22} \delta_{22}^2) - n_1 R_1 \delta_1^2 n_{21}^* R_{21} \delta_{21}^2}$$

It can be seen that weights are obtained from splitting variance of 1 into three components, with each being less than 1 and greater than 0, and each coefficient of $Z_k, \; k = 1, 21 \; or \; 22$ is the square root of corresponding weight. This is indeed very similar to variance spending method which dispenses variance into three independent test statistics. Let π_{21} be the rate of attrition/exclusion for placebo responders in Period 1 and π_{22} be the rate of attrition/exclusion of drug non-responders in Period 1. Thus, the expected sample size $n_{21}^* \; and \; n_{22}^*$ in Period 2 can be represented as a function of randomization ratio together with $\pi's$. That is: $n_{21}^* = \frac{n_1 r_1(1-\pi_{21})}{1+r_{21}}$ and n_{22}^*=$n_1(1-R_{22})(1-\pi_{22})$. With $\tau_{21}^2 = \frac{\delta_{21}^2}{\delta_1^2}$, $\tau_{22}^2 = \frac{\delta_{22}^2}{\delta_1^2}$, for given two-sided type I error α and type II error β, the required sample size for n_1 is: $n_1 = \frac{\left(\mathrm{z}_{1-\beta}+\mathrm{z}_{1-\frac{\alpha}{2}}\right)^2}{\delta_1^2 R_B}$, where $R_B = R_1 + \frac{R_{21}R_1(1-R_{21})(1-\pi_{21})}{1-R_1}\tau_{21}^2 + \; (1-R_{22})R_{22}(1-\pi_{22})\tau_{22}^2$. Enrichment of placebo non-responders alone (Liu et al. 2012) is a special case of the proposed method here. That is: without re-randomization of drug responders into Period 2, one now has $\lambda_1^* = \frac{n_1 R_1 \delta_1^2}{n_1 R_1 \delta_1^2 + n_{21}^* R_{21} \delta_{21}^2}$ and $n_1 = \frac{\left(\mathrm{z}_{1-\beta}+\mathrm{z}_{1-\frac{\alpha}{2}}\right)^2}{\delta_1^2 R_1 R_A}$ with $R_A = 1 + \frac{R_{21}(1-R_{21})(1-\pi_{21})}{1-R_{21}}\tau_{21}^2$. Because $R_B = \; R_A + (1 - R_{22})R_{22}(1-\pi_{22})\tau_{22}^2 > R_A$, sample size for enrichment of both Period 1 placebo non-responders and Period 1 drug responders can further decrease sample size and hence increase efficiency of the design compared to a SPD design with only re-randomizing Period 1 placebo non-responders into Period 2.

Section 6.3.2: Sample Size and Optimal Weight(s) Calculations

Optimal weight(s) and sample size are calculated for enrichment of placebo non-responders alone (Table 6.1) and for enrichment of both placebo non-responders and drug responders (Table 6.2) under a variety of scenarios.

Table 1 contains the results for enrichment of placebo non-responders alone, $r_1 = 2$ corresponding to 2:1 randomization ratio in Period 1, which is also proposed in various papers with SPD design to ensure that enough subjects can enter into Period 2. A special case of

$r_1 = 2$ shows that increase in sample size leads to higher power of the trial. Total sample size for Period 1 is 114 for power 0.8 while power is equal to 0.9 for sample size of 152 when $\delta_1 = \delta_{21} = 0.5$. With power of 0.8, sample size for total number of subjects in Period 1 decreases from 114 to 104 when having 20% increase in effect size (δ_{21}=0.6) in Period 2 from Period 1 (δ_1=0.5), as compared with the case with no change after enrichment (i.e., $\delta_1 = \delta_{21} = 0.5$*).* Similar trends also occur with other values of r_1. Considering varying value of r_1, one notices that weight $\lambda_{1,opt}$ decreases as r_1 increases. And the optimal weights for the listed scenarios are ranging from 0.6 to 0.8, consistent with published numbers in the literature.

Table 26(Tab. 6.1): Optimal rates and sample sizes for SPD

Table 6.1: For a SPD trial with enrichment of placebo non-responder, calculation of optimal $\boldsymbol{\lambda_{1,opt}}$ and sample size when $\boldsymbol{\alpha/2 = 0.025}$, $\boldsymbol{\beta = 0.1 \text{ } or \text{ } 0.2}$, $\boldsymbol{\delta_1 = 0.5}$, $\boldsymbol{\delta_{21} = 0.5 \text{ } or \text{ } 0.6}$, and $\boldsymbol{r_1 = 1.5, 1.7, 2.0, 2.2 \text{ } or \text{ } 2.5}$.

$1-\beta$	π_{21}	δ_1	δ_{21}	ε_{21}	r_1	$\lambda_{1,opt}$	n_1	$N_1 = \mathrm{n}_1 + \mathrm{n}_1 * \mathrm{r1}$
0.8	0.5	0.5	0.5	1.0	1.5	0.76	42	105
					1.7	0.75	40	108
					2.0	0.73	38	114
					2.2	0.71	37	117
					2.5	0.70	36	124
0.8	0.5	0.5	0.6	1.2	1.5	0.69	39	97
					1.7	0.67	37	99
					2.0	0.65	35	104
					2.2	0.63	34	107
					2.5	0.61	33	114
0.9	0.5	0.5	0.5	1.0	1.5	0.76	57	141
					1.7	0.75	54	145
					2.0	0.73	51	152
					2.2	0.71	49	156
					2.5	0.70	48	165
0.9	0.5	0.5	0.6	1.2	1.5	0.69	52	129
					1.7	0.67	50	133
					2.0	0.65	47	140
					2.2	0.63	45	144
					2.5	0.61	44	152

Table 6.2 repeats the calculations in Table 6.1 but with enrichment of both Period 1 placebo non-responders and Period 1 drug responders. Comparing with enrichment of placebo non-responders

alone (Table 6.1), in ESPD trials, total sample size decreases by 15%-20% as compared with cases in Table 6.1, resulting in more efficient designs. For $r_1 = 2$, all cases result in substantially saving in sample size as compared with respective cases in Table 6.1. Since both Period 1 placebo non-responders and Period 1 drug responders continue into Period 2 in ESPD trials, balanced randomization in Period 1 is more desirable and hence one could have r_1 around 1 rather than a number bigger than 1 as in Table 6.1. Note that in the proposed ESPDs, when $r_1 = 1$, equal effect size (i.e. $\delta_1 = \delta_{21} = \delta_{22} = 0.5$) and power 0.8, the required total sample size in Period 1 is 84; and as expected, sample size decreases to 79 when enrichment works and the effect size increases to 0.6 in Period 2 from being 0.5 in Period 1. From Table 6.2, it is also clear that optimal λ_1 is between 0.4 and 0.7, while λ_2 being a positive number less than 0.3. Compared to a simple parallel design, trials with SPD will save 30% in sample size (Liu et al. 2012) and further saving about 15%-20% in sample size can be achieved by ESPD compared to SPD trial.

Table 27(Tab. 6.2): Optimal rates and sample sizes for ESPD

Table 6.2: For an ESPD with both enrichment of placebo non-responders and drug responders, calculation of optimal $\lambda_{1,opt}$, $\lambda_{2,opt}$ and sample size when $\alpha/2 = 0.025$, $\beta = 0.1 \ or \ 0.2$, $\delta_1 = 0.5$, $\pi_{21} = \pi_{21} = 0.5, \delta_{21} = 0.5$, $\delta_{22} = 0.5 \ or \ 0.6$, and $r_1 = 0.5, 0.7, 1.0, 1.2, 1.5, 1.7, 2.2 \ or \ 2.5$.

$1-\beta$	π_{21}	π_{22}	δ_1	δ_{21}	δ_{22}	ε_{21}	ε_{22}	r_1	$\lambda_{1,opt}$	$\lambda_{2,opt}$	n_1	N_1
0.8	0.5	0.5	0.5	0.5	0.5	1	1	0.5	0.640	0.120	61	91
								0.7	0.660	0.140	51	86
								1.0	0.667	0.167	42	84
								1.2	0.665	0.182	39	85
								1.5	0.658	0.206	35	87
								1.7	0.651	0.220	33	88
								2	0.640	0.240	31	91
								2.2	0.632	0.253	29	93
								2.5	0.620	0.271	28	96
0.8	0.5	0.5	0.5	0.6	0.6	1.2	1.2	0.5	0.402	0.075	55	82
								0.7	0.421	0.089	47	79
								1.0	0.431	0.108	40	79
								1.2	0.433	0.119	36	79
								1.5	0.431	0.135	33	82
								1.7	0.428	0.144	31	83
								2	0.422	0.158	29	86

								2.2	0.418	0.167	28	88
								2.5	0.411	0.180	27	91
0.9	0.5	0.5	0.5	0.5	0.5	1	1	0.5	0.640	0.120	81	122
								0.7	0.660	0.140	68	115
								1.0	0.667	0.167	57	113
								1.2	0.665	0.183	52	113
								1.5	0.658	0.205	47	116
								1.7	0.651	0.220	44	118
								2	0.640	0.240	41	122
								2.2	0.632	0.253	39	124
								2.5	0.620	0.271	37	128
0.9	0.5	0.5	0.5	0.6	0.6	1.2	1.2	0.5	0.402	0.075	73	110
								0.7	0.421	0.089	62	106
								1.0	0.431	0.108	53	105
								1.2	0.433	0.119	49	106
								1.5	0.431	0.135	44	109
								1.7	0.428	0.144	42	112
								2	0.422	0.158	39	115
								2.2	0.418	0.167	37	118
								2.5	0.411	0.180	35	122

Section 6.4: Linear Combination Test in An Extended SPD with Binomial Data

Section 6.4.1: Preliminaries

For binary data collected from both periods, as shown in Table 6.3, there are four groups of patients across two periods: 1) patients who receive placebo in Period 1, are non-responders and re-randomized to receive placebo in Period 2 (PP), 2) patients who receive placebo in Period 1, are non-responders and re-randomized to receive drug in Period 2 (PD), 3) patients who receive drug in Period 1, are responders and then re-randomized to receive placebo in Period 2 (DP), and 4) patients who receive drug in Period 1, are responder and re-randomized to receive drug in Period 2 (DD).

Define $p_1 = P(drug\ response\ in\ Period\ 1)$, $q_1 = P(placebo\ response\ in\ Period\ 1)$, $p_{21} = P(drug\ response\ in\ Period\ 2|placebo\ non-responder\ in\ Period\ 1)$, $q_{21} = P(placebo\ response\ in\ Period\ 2|placebo\ non-responder\ in\ Period\ 1)$, $p_{22} = P(drug\ response\ in\ Period\ 2|drug\ responder\ in\ Period\ 1)$, $q_{22} = P(placebo\ response\ in\ Period\ 2|drug\ responder\ in\ Period\ 1)$.

Here, $p_{21}, q_{21,} p_{22}, and\ q_{22}$ are all conditional probabilities. Among PP subjects, n_{11} denotes the observed number of responders in Period 2; n_{12} denotes the observed number of subjects who are non-responders in both periods; n_{1A} is the observed number of subjects who are placebo responders in Period 1; n_1 is the total number of PP subjects and therefore $n_1 = n_{11} + n_{12} + n_{1A}$. Vector (n_{11}, n_{12}, n_{1A}) is multinomially distributed with $(n_1, (1-q_1)q_{21}$, (1-q_1)(1-q_{21}), q_1). n_{21} is, of the PD subjects, the observed number of subjects who are non-responders in Period 1 and responders in Period 2; n_{22} is, of the PD subjects, the observed number of subjects who are non-responders in both periods; n_{2A} is, of the PD subjects, the observed number of subjects who are placebo responders in Period 1; n_2 is the total number of PD subjects and $n_2 = n_{21} + n_{22} + n_{2A}$. Vector (n_{21}, n_{22}, n_{2A}) is multinomially distributed as $(n_2, (1-q_1)p_{21}$, (1-q_1)(1-p_{21}), q_1). n_{3B} is, of the DP subjects, the observed number of subjects who are drug non-responders in Period 1; n_{33} is, of the DP subjects, the observed number of subjects who are responders in both periods; n_{34} is, of the DP subjects, the observed number of subjects who are responders in Period 1 and non-responders in Period 2. $n_3 = n_{3B} + n_{33} + n_{34}$. Vector (n_{3B}, n_{33}, n_{34}) is multinomially distributed as $(n_3, (1-p), p_1 q_{22}, p_1(1-q_{22}))$. n_{4B} is, of the DD subjects, the observed number of subjects who are drug non-responders in Period 1; n_{43} is, of the DD subjects, the observed number of subjects who are responders in both periods; n_{44} is, of the DD subjects, the observed number of subjects who are responders in Period 1 and non-responders in Period 2. $n_4 = n_{4B} + n_{43} + n_{44}$. Vector (n_{4B}, n_{43}, n_{44}) is multinomially distributed as $(n_4, (1-p_1), p_1 p_{22}, p_1(1-p))$. The total sample size of the trial is n and n= $n_1 + n_2 + n_3 + n_4$. For sample size estimation and simulation of rejection probabilities, for the purpose of convenience, it is set to have $n_1 = n_2 = n_3 = n_4 = n/4$. Table 6.3 depicts the distribution of count data described above.

Table 28(Tab. 6.3): Extended sequential parallel design with binary data

Table 6.3: Extended sequential parallel design with binary data.

Treatment		Response					
Period 1	Period 2	Period 1	Period 2		Count		Probability
Placebo	Placebo (n_1)	No	Yes		n_{11}		(1-q_1) q_{21}
		No	No		n_{12}		(1-q_1)(1-q_{21})
		Yes	X		n_{1A}		q_1
Placebo	Drug (n_2)	No	Yes		n_{21}		(1-q_1) p_{21}
		No	No		n_{22}		(1-q_1)(1-p_{21})
		Yes	X		n_{2A}		q_1
Drug	Placebo n_3	No	X		n_{3B}		(1-p_1)
		Yes	Yes		n_{33}		p_1q_{22}
		Yes	No		n_{34}		$p_1(1-q_{22})$
Drug	Drug (n_4)	No	X		n_{4B}		(1-p_1)
		Yes	Yes		n_{43}		p_1p_{22}
		Yes	No		n_{44}		$p_1(1-p_{22})$

Section 6.4.2: Linear Combination Test

To test potential drug effect across two periods of the trial, we propose using maximum likelihood estimators from two periods; and then obtaining the linear combination of the two estimators, say h. Because estimated $\widehat{h}$ after plugging in maximum estimators is a function of all count vectors which have four different multinomial distributions, utilizing asymptotical normality of multinomial counts, delta method can be used to derive asymptotical variance of $\hat{h}$. The joint likelihood for observed data is defined as

$$\mathrm{L}={p_1}^{n_{33}+n_{34}+n_{43}+n_{44}}(1-p_1)^{n_{4B}+n_{3B}}{q_1}^{n_{1A}+n_{2A}}(1-q_1)^{n_{11}+n_{12}+n_{21}+n_{22}}{p_{21}}^{n_{21}}(1-p_{21})^{n_{22}}{q_{21}}^{n_{11}}(1-q_{21})^{n_{12}}{p_{22}}^{n_{43}}(1-p_{22})^{n_{44}}{q_{22}}^{n_{33}}(1-q_{22})^{n_{34}}$$

$$\log\mathrm{L}=(\mathrm{n}_{33}+\mathrm{n}_{34}+\mathrm{n}_{43}+\mathrm{n}_{44})\log(\mathrm{p}_1)+(\mathrm{n}_{4\mathrm{B}}+\mathrm{n}_{3\mathrm{B}})\log(1-\mathrm{p}_1)+(\mathrm{n}_{1\mathrm{A}}+\mathrm{n}_{2\mathrm{A}})\log(\mathrm{q}_1)+$$

$$(\mathrm{n}_{11}+\mathrm{n}_{12}+\mathrm{n}_{21}+\mathrm{n}_{22})\log(1-\mathrm{q}_1)+\mathrm{n}_{21}\log(\mathrm{p}_{21})\ +\mathrm{n}_{22}\log(1-\mathrm{p}_{21})+\mathrm{n}_{11}\log(\mathrm{q}_{21})+$$

$$\mathrm{n}_{12}\log((1-\mathrm{q}_{21})+\ \mathrm{n}_{43}\log(\mathrm{p}_{22})+\ \mathrm{n}_{44}\log(1-\mathrm{p}_{22})+\mathrm{n}_{33}\log(\mathrm{q}_{22})+\mathrm{n}_{34}\log(1-\mathrm{q}_{22})$$

$\hat{\mathrm{h}}=\mathrm{w}_1(\hat{\mathrm{p}}_1-\hat{\mathrm{q}}_1)\quad+\mathrm{w}_2(\hat{\mathrm{p}}_{21}-\hat{\mathrm{q}}_{21})+(1-\mathrm{w}_1-\mathrm{w}_2)(\hat{\mathrm{p}}_{22}-\hat{\mathrm{q}}_{22})$, where w_1 and w_2 are pre-specified weights. Under the situation of zero drug-placebo difference, $p_1=q_1, p_{21}=$

$q_{21,}\ p_{22} = q_{22,}$ h then equals 0. More effective a drug is, a bigger value h will become. The maximum likelihood estimate (MLE) can be solved by setting the first derivative of logL to 0. $\hat{p}_1 = \frac{n_{33}+n_{34}+n_{43}+n_{44}}{n_{33}+n_{34}+n_{43}+n_{44}+n_{4B}+n_{3B}}$, $\hat{q}_1 = \frac{n_{1A}+n_{2A}}{n_{11}+n_{12}+n_{21}+n_{22}+n_{1A}+n_{2A}}$, $\hat{p}_{21} = \frac{n_{21}}{n_{21}+n_{22}}$, $\hat{q}_{21} = \frac{n_{11}}{n_{11}+n_{12}}$, $\hat{p}_{22} = \frac{n_{43}}{n_{43}+n_{44}}$, $\hat{q}_{22} = \frac{n_{33}}{n_{33}+n_{34}}$. The maximum likelihood of h, $\hat{h}_{MLE}$, is obtained by substituting the MLEs into h function and the variance of $\hat{h}_{MLE}$ can be estimated using delta method. Define D^T= $[\frac{\partial \hat{h}}{\partial n_{11}}, \frac{\partial \hat{h}}{\partial n_{12}}, \frac{\partial \hat{h}}{\partial n_{1A}}, \frac{\partial \hat{h}}{\partial n_{21}}, \frac{\partial \hat{h}}{\partial n_{22}}, \frac{\partial \hat{h}}{\partial n_{2A}}, \frac{\partial \hat{h}}{\partial n_{3B}}, \frac{\partial \hat{h}}{\partial n_{33}}, \frac{\partial \hat{h}}{\partial n_{34}}, \frac{\partial \hat{h}}{\partial n_{4B}}, \frac{\partial \hat{h}}{\partial n_{43}}, \frac{\partial \hat{h}}{\partial n_{44}}]$ and define V=cov($[n_{11}\ n_{12}\ n_{1A}\ n_{21}\ n_{22}\ n_{2A}\ n_{3B}\ n_{33}\ n_{34}\ n_{4B} n_{43}\ n_{44}]^{\mathrm{T}}$). Then asymptotic Var($\hat{h}_{MLE}$)=$\hat{D}\hat{V}\hat{D}^T$. Since $n_{1,} n_{2,}\ n_{3,}$ $and\ n_{4,}$ are multinomially distributed and resulting from four count vectors of $(n_{11,}, n_{12,}, n_{1A})$, $(n_{21,}, n_{22,}, n_{2A})$, $(n_{3B,}, n_{33,}, n_{34})$, and $(n_{4B,}, n_{43,}, n_{44})$ respectively. For instance, Var(n_{11})= n_1(1-(1-q_1) q_{21}) (1-q_1) q_{21} and Cov(n_{11}, n_{12})= Cov(n_{12}, n_{11})=- n_1(1-q_1) q_{21}(1-q_1)(1-q_{21}). Similarly, all other variances and covariance can be easily derived. Thus, V=cov($[n_{11}\ n_{12}\ n_{1A}\ n_{21}\ n_{22}\ n_{2A}\ n_{3B}\ n_{33}\ n_{34}\ n_{4B} n_{43}\ n_{44}]^{\mathrm{T}}$), a 12X12 block diagonal matrix.

The resulting statistic is $T_{lc} = \frac{\hat{h}}{\sqrt{\hat{D}\hat{V}\hat{D}^T}}$, which converges to standard normal under null hypothesis. This shows that this linear combination test is a Wald test under null hypothesis of h=0.

Section 6.4.2: Sample Size Requirement and Simulated Rejecting Probabilities

Based on asymptotic normal of T_{lc}, sample size can be derived. Plugging in expected values of $n_{11}, n_{12,} n_{1A},\ n_{21}, n_{22,} n_{2A,}, n_{3B}, n_{33,} n_{34},\ n_{4B}, n_{43,}\ n_{44}$ into DVD^T and let $h = w_1(p_1 - q_1) + w_2(p_{21} - p_{21}) + (1 - w_1 - w_2)(p_{22} - q_{22})$, expected value of T_{lc} is $\frac{h}{\sqrt{DVD^T}}$. To achieve two-sided type I error of α and type II error of β, E(T_{lc})$|_{H_A}$=$z_{1-\alpha/2}$+$z_{1-\beta}$, where $z_{1-\alpha/2}$ is the

$\left(1-\frac{\alpha}{2}\right)th$ quintiles of the standard normal variable. Figure 2 shows the expected value of T_{lc} under alternative for different n. The horizontal dot-dashed line shows the value of $z_{1-\alpha/2}+z_{1-\beta}$ when $\alpha = 0.05$ and $\beta = 0.1$. At the point that the horizontal dot-dashed line intercepts, one draws a vertical line to intercept with the x-axis. The value at x-axis corresponds to the required sample size for a trial. For instance, the solid line is for drug-placebo difference being 0.1 for both periods and the required sample size to achieve power 0.9 is 620. The dashed line is for drug effect of 0.1 and 0.2 in Period 1 and Period 2 respectively and it requires n to be 252. When drug-placebo difference is 0.2 for both periods, it requires 113 for total n (dotted line in Figure 6.2). Comparing dashed line with dotted line, one can see clearly that sample size saves substantially when enrichment works in Period 2 (i.e., n = 620 versus n = 252).

Figure 18(Fig. 6.2): Graphic method for determining sample size

Figure 6.2: Graphic method for determining sample size. Expected value of T_{lc} under alternative hypothesis for different sample size at the beginning of Period 1 for $\mathbf{w_1 = 0.5}$ and $\mathbf{w_2 = 0.2}$. The solid line is for $\mathbf{p_1 = 0.7, q_1 = 0.6, p_{21} = 0.7,\ q_{21} = 0.6, p_{22} = 0.7, q_{22} = 0.6}$; The dashed line is for $\mathbf{p_1 = 0.7, q_1 = 0.6, p_{21} = 0.7, q_{21} = 0.5, p_{22} = 0.7, q_{22} = 0.5}$; the dotted line is for $\mathbf{p_1 = 0.7, q_1 = 0.5,\ p_{21} = 0.7, q_{21} = 0.5, p_{22} = 0.7, q_{22} = 0.5}$; and the horizontal dot-dashed line is the required expected mean under alternative hypothesis when 2-sided alpha is 0.05 and beta is 0.1.

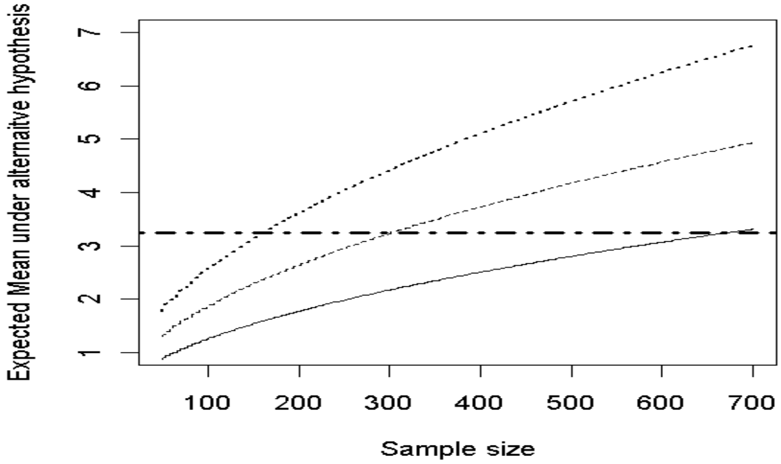


Table 6.4 shows the rejection error rates under null hypothesis for four scenarios of parameter profiles. Five cases of weight combinations are used. Based on explorations carried on bellow,

optimal weights for extended SPDs are for w_1 from 0.5-0.7 and w_1 from 0.15-0.25. 10000 simulation runs are used for all simulation experiments. It is clear that type I error rate is well controlled for all chosen parameters and weight combinations when sample size ranging from 50 to 1000. Note that the empirical type I error rates here are all subject to simulation errors.

Table 29(Tab. 6.4): Empirical one-sided type I error (X100)

Table 6.4: Empirical one-sided type I error (X100).

	n	$w_1 = 0.5$ $w_2 = 0.2$	$w_1 = 0.5$ $w_2 = 0.3$	$w_1 = 0.6$ $w_2 = 0.15$	$w_1 = 0.6$ $w_2 = 0.20$	$w_1 = 0.7$ $w_2 = 0.15$
$q_1 = 0.6$ $q_{21} = 0.4$ $q_{22} = 0.4$	50	2.91	3.42	3.23	3.25	3.20
	100	3.23	3.40	3.66	3.44	2.95
	150	3.13	3.20	3.05	2.87	2.94
	200	3.06	3.21	2.97	3.37	2.85
	300	3.27	3.41	2.92	2.87	2.68
	400	2.85	2.86	2.90	2.90	2.85
	500	2.99	2.91	3.05	2.58	3.20
	800	2.55	2.98	2.71	2.74	3.05
	1000	2.67	2.85	2.64	2.82	2.69
$q_1 = 0.5$ $q_{21} = 0.3$ $q_{22} = 0.3$	50	3.06	3.37	3.16	3.00	3.24
	100	3.57	3.67	3.23	3.60	2.87
	150	3.19	3.13	3.36	3.13	2.74
	200	3.20	3.54	3.15	2.92	2.77
	300	3.10	2.90	3.02	2.87	2.90
	400	3.07	3.35	2.77	2.90	2.56
	500	2.81	3.13	2.97	2.90	2.67
	800	3.07	2.42	2.87	2.68	2.53
	1000	2.75	2.91	2.96	2.43	2.49
$q_1 = 0.4$ $q_{21} = 0.2$ $q_{22} = 0.2$	50	3.61	3.47	3.39	3.21	3.05
	100	3.37	3.30	3.28	2.95	2.64
	150	3.45	2.93	3.46	3.04	2.78
	200	3.30	3.06	3.05	3.34	2.65
	300	3.28	2.90	3.05	3.00	2.67
	400	2.83	3.03	3.12	2.79	2.68
	500	2.68	3.10	2.76	2.84	2.67
	800	3.02	2.83	2.72	2.74	2.69
	1000	2.75	3.01	2.99	2.82	2.71

Table 6.5 contains calculation of the required sample size based on the method described in Figure 6.2 for various parameter-weight combinations. After obtaining sample sizes, simulations are conducted with 10000 simulation runs for each scenario. There are 3 sets of simulations. Case A: drug-placebo difference $(p_r - q_r)$, where index r =1, 21, or 22 all being 0.1 in both periods, which includes three subtypes with the probability of being a placebo responder being

0.6, 0.5 and 0.4 respectively. Case B: drug-placebo difference being 0.1 and 0.2 for Period 1 and Period 2, respectively, which includes three subtypes with $q_1 = 0.6$ and $q_{21} = q_{22} = 0.5$; $q_1 = 0.5$ and $q_{21} = q_{22} = 0.4$ and $q_1 = 0.4$ and $q_{21} = q_{22} = 0.3$, respectively. Case C: drug-placebo difference being 0.2 in both periods, which contains three subtypes with $q_r = 0.5, 0.4, 0.3, r = 1,21,22$, respectively.

If drug effect is 0.1 in both periods (Case A), 0.1 in Period 1 and 0.2 in Period 2 (Case B) and drug effect is 0.2 in both periods (Case C), it is clear that the required sample size decreases from Case A to Case C (Table 6.5). It confirms that it is easier to detect drug superiority when either enrichment works (Case B versus Case A) and/or drug effect size increases (Case C versus Case B).

In all cases of simulations, empirical powers are always smaller than the target power of 0.9 used for calculating sample size. However, the extent of power decrease shows interesting patterns. In Case A, when drug-placebo is equal to 0.1, the required sample size is high, but the simulated power is only 3-4% less than the design value 90%; in Case B, when drug-placebo difference increases from 0.1 in Period 1 to 0.2 in Period 2, the simulated power was 5-8% less than the design value 90%; in Case C, when drug-placebo difference is 0.2 for both periods, the required sample size is only a little more than 100, but the simulated power is 15-18% less than the design value 90%. This is an alert to us because we normally use calculated sample size directly to plan a trial, or just increase sample size by 10% to ensure power. But our examinations on empirical powers in extended SPD trials tell us that 10% increase from the calculated sample size based on asymptotic normality as suggested in Liu et al. (2012) can’t always guarantee enough power in real practices. And the required sample size in real practices may depend on the particular parameter profile of interest and may require extensive simulation explorations prior to trial start rather than lazily using the calculated sample size based on asymptotic normality.

The results of the simulations show the impacts of pre-specified weights on trial powers. For instance, among five scenarios with $p_1 = 0.7, q_1 = 0.6,\ p_{21} = 0.7, q_{21} = 0.6, p_{22} = 0.7,\ q_{22} = 0.6$, the highest sample size is 721 occurring at $w_1 = 0.5$ and $w_2 = 0.3$ while the lowest is 157 (a decrease of 564 from 721) occurring when $w_1 = 0.6$ and $w_2 = 0.15$. However, no specific rules can be summarized here. One also notices that sample size has a small variation among the explored scenarios in Case C when having a relatively large drug-effect of 0.2 in both periods. Tables 6.4 – 6.5 show that sample sizes calculated using asymptotic properties of linear combination test are good enough for conducting clinical trials. However, it would be better to conduct extensive simulations for various parameter profiles of interest prior to trial start since there is a difference in extent of power deduction probably caused by insufficiency in asymptotic normality.

Table 30(Tab. 6.5): Required sample size and empirical power(X100) simulation

Table 6.5: Required sample size and empirical power(X100) simulation.

		$w_1 = 0.5$ $w_2 = 0.2$		$w_1 = 0.5$ $w_2 = 0.3$		$w_1 = 0.6$ $w_2 = 0.15$		$w_1 = 0.6$ $w_2 = 0.20$		$w_1 = 0.7$ $w_2 = 0.15$	
		n	power	n	power	n	power	n	power	n	power
Case A	$p_1 = 0.7, q_1 = 0.6$ $p_{21} = 0.7, q_{21} = 0.6$ $p_{22} = 0.7, q_{22} = 0.6$	620	85.7	721	86.2	564	84.9	588	85.8	580	85.9
	$p_1 = 0.6, q_1 = 0.5$ $p_{21} = 0.6, q_{21} = 0.5$ $p_{22} = 0.6, q_{22} = 0.5$	681	86.3	716	86.4	628	86.4	625	85.9	625	87.1
	$p_1 = 0.5, q_1 = 0.4$ $p_{21} = 0.5, q_{21} = 0.4$ $p_{22} = 0.5, q_{22} = 0.4$	716	85.8	681	86.1	657	86.0	625	85.8	628	86.6
Case B	$p_1 = 0.7, q_1 = 0.6$ $p_{21} = 0.7, q_{21} = 0.5$ $p_{22} = 0.7, q_{22} = 0.5$	252	80.3	297	81.8	268	81.7	281	81.5	324	83.3
	$p_1 = 0.6, q_1 = 0.5$ $p_{21} = 0.6, q_{21} = 0.4$ $p_{22} = 0.6, q_{22} = 0.4$	273	81.8	284	81.2	292	82.4	292	82.3	384	84.5
	$p_1 = 0.5, q_1 = 0.4$ $p_{21} = 0.5, q_{21} = 0.3$ $p_{22} = 0.5, q_{22} = 0.3$	276	81.5	265	81.6	300	81.9	284	82.9	345	84.7
Case C	$p_1 = 0.7, q_1 = 0.5$ $p_{21} = 0.7, q_{21} = 0.5$ $p_{22} = 0.7, q_{22} = 0.5$	113	72.6	124	72.9	100	71.2	105	72.2	105	72.7

	$p_1 = 0.6, q_1 = 0.4$ $p_{21} = 0.6, q_{21} = 0.4$ $p_{22} = 0.6, q_{22} = 0.4$	124	73.2	124	74.4	113	72.5	108	72.6	113	74.5
	$p_1 = 0.5, q_1 = 0.3$ $p_{21} = 0.5, q_{21} = 0.3$ $p_{22} = 0.5, q_{22} = 0.3$	124	72.7	113	72.2	113	72.2	105	72.4	105	72.7

Section 6.5: Discussions

In this article, we introduce an ESPD. In this design, placebo responders and drug non-responders during period 1 are re-randomized to receive placebo or drug during period 2 of the trial. The proposed statistics to test superiority of drug against placebo is the optimal weight Z test for normal data, which requires deriving optimal weight upfront. After evaluating clinical outcomes from two periods, weight Z test with optimal weights will be used to combine information from three cohorts, one from Period 1 and two from Period 2. This is different from the design suggested by Fava et al. (2003) which does not have the second randomization. It is also different from design considered by Chen et al. (2011) and Liu et al. (2012) where only placebo non-responders during Period 1 are re-randomized prior to period 2. Since we extend Liu et al. (2012) to further include Period 1 drug responders into Period 2, other related discussions in Liu et al. (2012) such as controlling baseline variables, multiplicity issue, using trend test in certain contexts and so on can also be utilized here. For binary data, linear combination test for ESPD trials is proposed in Section 4. Sample size can be planned using a graphic method. Simulations are done to evaluate type I error rate controlling and power achievement in ESPD and it is suggested that it is very important to conduct extensive simulations prior to trial start in order to extensively exam trial operational characteristics.

Reference

Chen, Y. F., Yang, Y., Hung, H. M. J., Wang, S. J. (2011). Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials* 32:592-604.

Cui, L., Hung, H. M. J., Wang, S. J.(1999). Modification of sample size in group sequential clinical trials. *Biometrics* 55:853-857.

Fava, M., Evins, A. E., Dorer, D. J., Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics* 72:115-127.

Habermann, T. M., Weller, E. A., Morrison, V. A., et al. (2006). Rituximab-CHOP versus CHOP alone or with maintenance rituximab in older patients with diffuse large B-cell lymphoma. *Journal of Clinical Oncology* 24:3121-3127.

Heyn, R. M., Joo, P., Karon, M., et al.(1974). BCG in the treatment of acute lymphocytic leukemia. *Blood* 46:431-442.

Khin, N.A., Chen, Y.F., Yang, Y., Yang, P., Laughren, T.P. Exploratory analyses of efficacy data from major depressive disorder trials submitted to the U.S. Food and Drug Administration in support of new drug applications. *J Clin Psychiatry*. 2011;72:464–472.

Mills, E. J., Kelly, S., Wu, P., Guyatt, G. H., (2007). Epidemiology and reporting of randomized trials employing rerandomization of patient groups: a systematic survey. *Contemporary Clinical Trials* 28:268-275.

Tamura, R., Huang, X., (2007). An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clinical Trials* 4:309-317.

Liu, Q., Lim, P., Singh, J., Lewin, D., Schwarb, B., Kent, J., (2012). Doubly randomization delayed start design for enrichment studies with responders or non-responders. *Journal of Biopharmaceutical Statistics* 22:4, 737-757.

Chapter 7

Covariance and Variance Evaluations of Two Estimators for Drug-placebo Difference in a Trial with Sequential Parallel Design

(accepted for publication in Jun 2016 by *Communication in Statistics: theory and method*)

Abstract: Chen et al. [Contemp. Clin. Trials, 32: 592-604 (2011)] heuristically proved that the covariance of two estimators is zero assuming equal correlation coefficients. In this article, above covariance is re-derived without any strong assumption in equality between two correlation coefficients. Under rigorous analytic derivations plus assuming number of subjects continuing into Period 2 is a random variable, covariance is re-confirmed to be zero for both normal and binomial data.

Keywords: Placebo Effect; Sequential Parallel Design; Drug-placebo Difference; Seemly Unrelated Regression.

Section 7.1: Introduction

In randomized double-blind clinical trials, subjects are randomized to receive either drug or placebo where the assigned treatment is unknown to both patients and investigators. By doing this, the drug-placebo difference on the endpoint will demonstrate the drug effect on patients if there is no placebo effect, since randomization has balanced out baseline covariates between drug and placebo groups and blinding can hopefully eliminate positive expectancy towards study drug during the trial. However, if the placebo response is relatively high in the trial, this drug-placebo difference decreases, which may result in the failure of detecting treatment effect. Adding a placebo lead-in period prior to randomization is the most conventional method to reduce placebo response. After the lead-in period, only placebo non-responders (based on predefined criteria/criterion) are randomized into the double-blind period where the drug-placebo difference is measured. Among 86 major depressive disorder (MDD) trials, least-squared mean change from baseline to endpoint for the Hamilton Rating Scale for Depression (HAMD) for placebo-treated subjects in thirty trials without the placebo lead-in period was -9.24 (SD=1.87), while for the two other types (differentiated by criterion for placebo responder) of trials with a placebo lead-in period it was -7.88 (SD=2.12) and -7.56 (SD=1.80) (Walsh et al. 2002).

The conventional parallel group has only one treatment period, whereas, Fava's sequential parallel design (Fava et al. 2003) has two treatment periods with Period 2 consisting of only placebo non-responders from Period 1. In Period 2, subjects either continue on placebo or receive treatment. At the end of the trial, inference on the drug-placebo difference for all subjects randomized in Period 1 ($\hat{\delta}_1$) and inference on the drug-placebo difference in Period 2 ($\hat{\delta}_2$) for Period 1 placebo non-responders is combined. The null hypothesis is H_0: δ_1= δ_2= 0, the alternative hypothesis is $H_A : \delta_1 > 0$ or $\delta_2 > 0$ and the combined estimator is $w\hat{\delta}_1 + (1-w)\hat{\delta}_2$. The sequential parallel design (SPD) is more efficient than the traditional parallel group design (1): $\hat{\delta}_2$ is estimated from Period 1 placebo non-responders, which is normally bigger than $\hat{\delta}_1$, and (2): Period 1 placebo non-responders contribute twice in testing $\hat{\delta}_1$ and $\hat{\delta}_2$, resulting in a larger 'effective' sample size than that of utilizing data collected from Period 1 only, and hence increases power.

To implement a SPD trial with continuous endpoints, Tamura and Huang (2007) proposed seemly unrelated regression (SUR). By stacking continuous data from two periods together, SUR simultaneously estimate the variance-covariance matrix and parameters of interests, and then constructs a test statistic based on the combined estimator and its variance. That is: $w^2\mathrm{Var}(\hat{\delta}_1) + 2w(1-w)\mathrm{cov}(\hat{\delta}_1, \hat{\delta}_2) + (1-w)^2\mathrm{Var}(\hat{\delta}_2)$, or $w^2\hat{\sigma}_{11} + 2w(1-w)\hat{\sigma}_{12} + (1-w)^2\hat{\sigma}_{22}$. The data from two periods can be expressed via a linear relationship: $Y_i = K_i\delta_i + \epsilon_i$, i=1,2, where Y_i is a vector of a continuous endpoint from the ith period, and Z_i is the design matrix of the ith period, assuming there is only one independent variable (i.e., treatment arm) in linear equation. K_i is either 1 for drug and 0 for placebo. The coefficient for K_1 is δ_1 and the coefficient for K_2 is δ_2. The size of Y_1 is the number of subjects in Period 1, and size of Y_2 is the number of placebo non-responders from Period 1 who continue into

Period 2. ϵ_1 is error term for the 1st regression and independently distributed with mean 0 and variance of σ_{11}^2 for every subject in Period 1; ϵ_2 is error term for the 2nd regression and independently distributed with mean 0 and variance of σ_{22}^2 for every subject in Period 2; and the covariance $\sigma_{12}(\text{or } \sigma_{21})$ for endpoints at Period 1 and Period 2 only for subjects who are placebo non-responders at the end of Period 1 and continue into Period 2. To estimate both δ_1 and δ_2, two linear equations are stacked to become a single linear model form of: $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \begin{bmatrix} K_1 & 0 \\ 0 & K_1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$and the within patient residual vector has a variance covariance matrix of: $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$. In the stacked linear model, there are three parameters of σ_{11}, σ_{22} and $\sigma_{12}(\sigma_{21})$ in Σ to be estimated from the data using ordinary least squares residuals, and then the coefficient vector of $\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$ will be obtained once the response vector, design matrix and Σ are known. When the sample size for both periods are large enough, $\hat{\Sigma}$ will be consistent.

At the beginning of Period 2 of an SPD trial, placebo non-responders can be re-randomized. For an SPD trial, the estimate for each period is used to evaluate drug-placebo difference. There are several methods to combine the evidences from two periods. When the Wald-test is used, the variance of the weighted estimators which is the key for hypothesis testing consists of calculating the intra-variability between the endpoints from two periods (i.e., covariance) and the variance of two estimators separately, with the latter being much easier to derive. If the covariance equal to zero, the complexity of the test in an SPD trial will be much reduced. In Chen et al. (2011), covariance of $\hat{\delta}_1$and $\hat{\delta}_2$ was further investigated and was shown to be zero for normal data. In their derivation, the sample size for Period 2 is a fixed number. This is a questionable assumption because being a placebo responder or a non-responder is a random variable and hence the number of placebo non-responders to enter Period 2 will also be a random variable. Furthermore, the equality assumption in the two correlation coefficients is also questionable.

To relax the limitations in the derivation of covariance, we derive the covariance between the two estimators for the scenario with normal endpoints in both Period 1 and Period 2 (i.e., normal-normal) and binomial-binomial in Sections 7.2 and 7.3, respectively. Section 7.2.1 lays out the proof structure for the normal-normal case; Section 7.2.2 revisits the sample size derivation under the assumption of the covariance being zero plus the assumption that the number of subjects continuing into Period 2 is a random variable; Section 7.2.3 performs simulation exercises assessing type I error rate and power under the conditional independence assumption; and Section 7.2.4 examines possible violations of the proposed independence assumption in Section 7.2.3. Section 7.3 repeats steps in Sections 7.2.1 – 7.2.3 but for binomial-binomial data, without conducting simulations under dependence structure because we lack a clear understanding on how binomial endpoints from the two periods are correlated in practice. In the end, Section 7.4 concludes this paper with discussions and further research directions hinted by research results here.

Section 7.2: Normal -Normal Data

Section 7.2.1: Covariance for $\hat{\delta}_1$ and $\hat{\delta}_2$, Re-examination

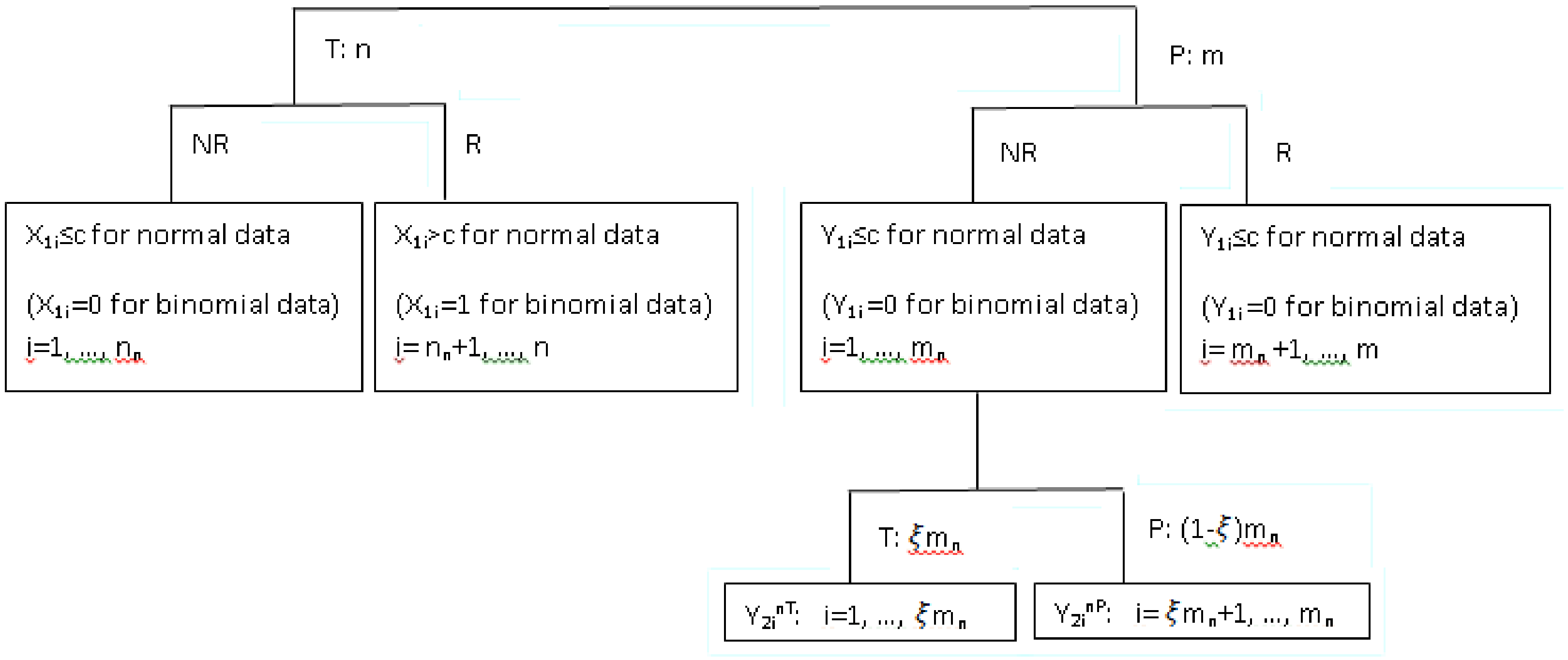

T: n
P: m
NR
R
NR
R
X_{1i}≤c for normal data
(X_{1i}=0 for binomial data)
i=1, ..., n_n
X_{1i}>c for normal data
(X_{1i}=1 for binomial data)
i= n_n+1, ..., n
Y_{1i}≤c for normal data
(Y_{1i}=0 for binomial data)
i=1, ..., m_n
Y_{1i}≤c for normal data
(Y_{1i}=0 for binomial data)
i= m_n +1, ..., m
T: ξm_n
P: (1-ξ)m_n
Y_{2i}^{nT}: i=1, ..., ξm_n
Y_{2i}^{nP}: i=ξm_n+1, ..., m_n

Figure 19(Fig. 7.1): A SPD trial

Figure 7.1: A SPD trial. NR and R denotes non-responders ($X_{1i} \leq$ c for normal data X_{1i} =0 for binomial data) and responders (X_{1i} > c for normal data X_{1i} = 1 for binomial data I = n_n +1, …, n). Similar definitions are defined for subjects in the treatment group. T and P denote treatment and placebo group respectively in both periods.

Suppose there are n subjects to be treated in Period1 by study drug and the corresponding endpoint, X_{1i}, i = 1,…, n, is normally distributed with mean μ_{T1} and variance σ_{T1}^2 at the end of Period 1, resulting in n_n non-responders with $X_{1i} \leq c$ while n- n_n responders with endpoint realized with a value greater than the threshold value c. In the meantime, there are m subjects to be treated in Period 1 by placebo and corresponding endpoint, Y_{1i}, i = 1,…, m, is normally distributed with mean μ_{P1} and variance σ_{P1}^2, resulting in m_n non-responders with $Y_{1i} \leq c$ while m- m_n responders with $Y_{1i} > c$. Unlike subjects in the treatment group, the placebo non-responders are enrolled in Period 2 for further assessment of the drug-placebo difference. Period 1 placebo non-responders who are on study drug in Period 2 will have endpoint, $Y_{2i}^{nT} | Y_{1i} \leq c$, i = 1,…, ξm_n, normally distributed with mean μ_{nT} and variance σ_{nT}^2, with ξ as the proportion of Period 1 non-responders being treated with study drug in Period 2. Similarly, non-responders treated with placebo in Period 2 will have endpoint, $Y_{2i}^{nP} | Y1i \leq c$, , ξm_n+1, …, m_n, normally distributed with mean μ_{nP} and variance σ_{nP}^2.

That is: $X_{1i} \sim \text{Normal}\left(\mu_{T1},\ \sigma_{T1}^2\right)$, i = 1,…, n; $Y_{1i} \sim \text{Normal}\left(\mu_{P1},\ \sigma_{P1}^2\right)$, $i = 1, \dots, m$

$Y_{2i}^{nT} | Y1i \leq c \sim \text{Normal}\left(\mu_{nT},\ \sigma_{nT}^2\right)$, i = 1,…, ξm_n ; $Y_{2i}^{nP} | Y1i \leq c \sim \text{Normal}\left(\mu_{nP},\ \sigma_{nP}^2\right)$, $i = \xi m_n + 1, \dots, m_n$

So the estimators of drug-placebo difference at Period 1and Period 2 respectively, are as follows:

$$\hat{\delta}_1 = \hat{\mu}_{T1} - \hat{\mu}_{P1} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} - \frac{1}{m}\sum_{i=1}^{m} Y_{1i}$$

$$\hat{\delta}_2 = \hat{\mu}_{nT} - \hat{\mu}_{nP} = \frac{1}{\xi m_n}\sum_{i=1}^{\xi m_n} Y_{2i}^{nT} - \frac{1}{(1-\xi) m_n}\sum_{i=\xi m_n+1}^{m_n} Y_{2i}^{nP}$$

$$cov\left(\hat{\delta}_1, \hat{\delta}_2\right) = cov(\hat{\mu}_{T1} - \hat{\mu}_{P1}, \hat{\mu}_{nT} - \hat{\mu}_{nP})$$

$$= cov(\hat{\mu}_{T1}, \hat{\mu}_{nT}) - cov(\hat{\mu}_{T1}, \hat{\mu}_{nP}) - cov(\hat{\mu}_{P1}, \hat{\mu}_{nT}) + cov(\hat{\mu}_{P1}, \hat{\mu}_{nP})$$

=0 per proof in Appendix 7.1.

Appendix 7.1 proves zero covariance for the normal-normal case. The covariance of Period 1 treatment-placebo difference of $\hat{\delta}_1$ and Period 2 treatment-placebo difference of $\hat{\delta}_2$ can be decomposed into four parts, in which $\mathrm{cov}(\hat{\mu}_{T1}, \hat{\mu}_{nT})$ and $\mathrm{cov}(\hat{\mu}_{T1}, \hat{\mu}_{nP})$ are both equal to zero because two estimators are drawn from different cohorts of subjects. Non-zero terms $\mathrm{cov}(\hat{\mu}_{P1}, \hat{\mu}_{nT})$ and $\mathrm{cov}(\hat{\mu}_{P1}, \hat{\mu}_{nP})$ are then calculated using the 'law of total covariance' so that the covariance is equal to sum of the expected covariance and the covariance of expectations, where the variable to be conditioned upon is the random variable of placebo non-responders (i.e., $I(Y_{1i} > c), i = 1, \dots, m)$ at the end of Period 1. For instance, after conditioning upon $I(Y_{1i} > c)$, $cov(\hat{\mu}_{P1}, \hat{\mu}_{nT})$ calculation becomes the expectation of conditional covariance plus the covariance of two conditional variables. That is: $cov(\hat{\mu}_{P1}, \hat{\mu}_{nP}) = E[cov(\hat{\mu}_{P1}, \hat{\mu}_{nP} | I(Y_{1i} > c), i = 1, \dots, m)] + cov\big(E(\hat{\mu}_{P1} | I(Y_{1i} > c), i = 1, \dots, m), E(\hat{\mu}_{nP} | I(Y_{1i} > c), i = 1, \dots, m)\big) = \mathcal{A} + \mathcal{B}$. Similarly, $cov(\hat{\mu}_{P1}, \hat{\mu}_{nT}) = \mathcal{A}' + \mathcal{B}'$. Hence $cov(\hat{\delta}_1, \hat{\delta}_2) = cov(\hat{\mu}_{P1}, \hat{\mu}_{nP}) - cov(\hat{\mu}_{P1}, \hat{\mu}_{nT})$ $= \mathcal{A} + \mathcal{B} - \{ \mathcal{A}' + \mathcal{B}' \}$.

$\mathcal{A} = E[cov(\hat{\mu}_{P1}, \hat{\mu}_{nP} | I(Y_{1i} > c), i = 1, \dots, m)]$ and the inner part under its expectation is the covariance of two conditional random variables, where $cov(\hat{\mu}_{P1}, \hat{\mu}_{nP} | I(Y_{1i} > c), i = 1, \dots, m)$ can be further decomposed into four expectations of the product of two quantities, which is either a conditional random variable or an expectation of a conditional random variable. Therefore, one has $\mathrm{cov}(\hat{\mu}_{P1}, \hat{\mu}_{nP} | \mathrm{I(Y1i>c)}, i = 1, \dots, m) = A - B - C + D$ with $\mathcal{A} = E(A) - E(B) - E(C) + E(D)$ (Appendix 7.1). Terms A, B, C and D are then respectively calculated for $\mathcal{A}$ and $\mathcal{A}'$ and simplified with help of the quantities of the mean and the variance of truncated normal random variables of $Y_{1i} | Y_{1i} \leq c$ and $Y_{1i} | \mathrm{Y1i} > c, i = 1, \dots, m$. When all terms are combined together, $\mathcal{A} - \mathcal{A}'$ is shown to be zero and with the help of the 'law of total expectation', which states that the expected value of the conditional expected value of R given S is the same as the expected value of R. Besides, both $\mathcal{B}$ and $\mathcal{B}'$ are shown to be zero per calculation. In summary, $\mathrm{cov}(\hat{\delta}_1, \hat{\delta}_2)$ is proved to be zero in an SPD trial with normal-normal data.

To simplify the understanding of this tedious proof in Appendix 7.1, a schematic is shown in Illustration 7.1 in Appendix 7.1, in which Step I makes use of the 'law of total covariance' and Step II utilizes the 'law of total expectation' when calculating $\mathrm{E(H), E(G), E(I)}$ and $\mathrm{E(J)}$. As pointed out by the reviewer, some people are not familiar with the term "conditional random variable" because a random variable is just a random variable and conditioning is for the purpose of calculating distribution property such as conditional expectations. We totally agree with these comments and also agree that the purpose of using conditional random variable in this paper is to help with the proof as what was done in deriving variance decomposition formula (or law of total variance) in probability theory.

Next, let's return to the proof of zero covariance between $\hat{\delta}_1$ and $\hat{\delta}_2$ by Chen and et al. (2011) and see how it differs from the proposed method here. From Chen and et al. (2011), the proof is re-written using notations in this paper as follows:

$$cov(\hat{\delta}_1, \hat{\delta}_2) = cov\left(\frac{1}{n}\sum_{i=1}^{n} X_{1i} - \frac{1}{m}\sum_{i=1}^{m} Y_{1i}\,, \frac{1}{\xi \mathrm{m_n}}\sum_{i=1}^{\xi \mathrm{m_n}} \mathrm{Y_{2i}^{nT}} - \frac{1}{(1-\xi)\mathrm{m_n}}\sum_{i=\xi \mathrm{m_n}+1}^{\mathrm{m_n}} \mathrm{Y_{2i}^{nP}}\right)$$

$$= cov\left(\frac{1}{n}\sum_{i=1}^{n} X_{1i}\,, \frac{1}{\xi \mathrm{m_n}}\sum_{i=1}^{\xi \mathrm{m_n}} \mathrm{Y_{2i}^{nT}}\right) - cov\left(\frac{1}{n}\sum_{i=1}^{n} X_{1i}\,, \frac{1}{(1-\xi)\mathrm{m_n}}\sum_{i=\xi \mathrm{m_n}+1}^{\mathrm{m_n}} \mathrm{Y_{2i}^{nP}}\right)$$

$$-cov\left(\frac{1}{m}\sum_{i=1}^{m} Y_{1i}\,, \frac{1}{\xi \mathrm{m_n}}\sum_{i=1}^{\xi \mathrm{m_n}} \mathrm{Y_{2i}^{nT}}\right) + cov\left(\frac{1}{m}\sum_{i=1}^{m} Y_{1i}\,, \frac{1}{(1-\xi)\mathrm{m_n}}\sum_{i=\xi \mathrm{m_n}+1}^{\mathrm{m_n}} \mathrm{Y_{2i}^{nP}}\right)$$

$$= 0 - 0 - cov\left(\frac{1}{m}\sum_{i=1}^{m} Y_{1i}\,, \frac{1}{\xi \mathrm{m_n}}\sum_{i=1}^{\xi \mathrm{m_n}} \mathrm{Y_{2i}^{nT}}\right) + cov\left(\frac{1}{m}\sum_{i=1}^{m} Y_{1i}\,, \frac{1}{(1-\xi)\mathrm{m_n}}\sum_{i=\xi \mathrm{m_n}+1}^{\mathrm{m_n}} \mathrm{Y_{2i}^{nP}}\right)$$

$$= -\frac{1}{m} * \frac{1}{\xi \mathrm{m_n}} * \xi \mathrm{m_n} * \rho\left(Y_{1i}, \mathrm{Y_{2i}^{nT}}\right) * \sigma_{\mathrm{P1}} * \sigma_{\mathrm{nT}} + \frac{1}{m} * \frac{1}{(1-\xi)\mathrm{m_n}} * (1-\xi)\mathrm{m_n} * \rho\left(Y_{1i}, \mathrm{Y_{2i}^{nP}}\right) * \sigma_{\mathrm{P1}} *$$

$$\sigma_{\mathrm{nP}} = -\frac{1}{m} * \rho\left(Y_{1i}, \mathrm{Y_{2i}^{nT}}\right) * \sigma_{\mathrm{P1}} * \sigma_{\mathrm{nT}} + \frac{1}{m} * \rho\left(Y_{1i}, \mathrm{Y_{2i}^{nP}}\right) * \sigma_{\mathrm{P1}} * \sigma_{\mathrm{nP}} = 0$$

The above derivation assumes $\rho\left(Y_{1i}, Y_{2i}^{nT}\right) = \rho\left(Y_{1i}, Y_{2i}^{nP}\right)$ as well as $\sigma_{nT} = \sigma_{nP}$, and also treats m_n as a constant. These questionable assumptions are no longer required in the proposed method here. However, it might be worthwhile to explain why zero covariance can be obtained when equal correlation (i.e., $\rho\left(Y_{1i}, Y_{2i}^{nT}\right) = \rho\left(Y_{1i}, Y_{2i}^{nP}\right)$) is removed heuristically besides using lengthy mathematical calculations. From our perspective, the most reasonable answer for this may be the stipulation of conditional independence between endpoints between two periods. That is, given normally distributed with mean μ_{nT} and variance σ_{nT}^2 for $Y_{2i}^{nT} | Y1i \leq c$ (or μ_{nP} and variance σ_{nP}^2 for $Y_{2i}^{nP} | Y_{1i} \leq c$), it is said that the Period 1 endpoint is independent of the Y_{1i} because the Period 2 endpoint is not a function of the realization of the Period 1 variable. The impact of this assumption on proposed method will be assessed below in Section 7.2.4.

Section 7.2.2: Sample Size Derivation and A Hypothetical Trial Example

After evaluating and re-confirming the zero covariance in Section 7.2.1 when the endpoints in Period 1 and Period 2 are both normal, re-examination of the variance of the weighted test statistic for an SPD trial will be done in this section. In Chen et al. (2011), the estimated rate of being a placebo non-responder at the end of Period 1 is used in the variance equation. However, with a pre-defined distribution for Period 1 data, the expected rate of being a placebo non-responder at end of Period 1 can be calculated and used for sample size calculation. For normal data, the probability of being a placebo non-responder, that is $Y_{1i} \leq c$, is $\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)$ with c as the cutoff point for being a responder. In the case of a binomial endpoint, the probability of being a placebo non-responder is $1 - P_P(r_1)$. The allocation ratio for placebo and treatment is $b{:}(1-b)$ in Period 1 and then equal allocation between two groups (i.e., 0.5:0.5) in Period 2. $b = 0.66$ is used for the sample size calculation in order to ensure more subjects to be randomized into the placebo group in Period 1.

With Z as a standard normal random variable and standardized weighted-z test of $T=\frac{w\hat{\delta}_1+(1-w)\hat{\delta}_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}$, the probability of rejecting null ($\delta_1 = \delta_2 = 0$) when alternative hypothesis is true ($\delta_1, \delta_2 > 0$) is:

$$P_{HA}\left(T > Z_{1-\frac{\alpha}{2}}\right) = P_{HA}\left(T < -z_{1-\frac{\alpha}{2}}\right) = P_{HA}\left(\frac{w\hat{\delta}_1+(1-w)\hat{\delta}_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}} < -z_{1-\frac{\alpha}{2}}\right)$$

$$=P_{HA}\left(\frac{w\hat{\delta}_1+(1-w)\hat{\delta}_2-(w\delta_1+(1-w)\delta_2)}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}} < -z_{1-\frac{\alpha}{2}} - \frac{w\delta_1+(1-w)\delta_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}\right)=P_{H0}\left(Z < -z_{1-\frac{\alpha}{2}} - \frac{w\delta_1+(1-w)\delta_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{w\delta_1+(1-w)\delta_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}\right),$$

where α is the type I error rate for this two-sided hypothesis test.

$\therefore\ z_{1-\beta} + z_{1-\frac{\alpha}{2}} = -\frac{w\delta_1+(1-w)\delta_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}$, where β is type II error to ensure probability of rejecting null when alternative hypothesis is true.

Thus, $w^2Var(\hat{\delta}_1) + (1-w)^2Var(\hat{\delta}_2) = (\frac{w\delta_1+(1-w)\delta_2}{z_{1-\beta}+z_{1-\frac{\alpha}{2}}})^2$

For normal data, $Var(\hat{\delta}_1) = \frac{\sigma_{T1}^2}{n} + \frac{\sigma_{P1}^2}{m} = \frac{\sigma_{T1}^2}{N(1-b)} + \frac{\sigma_{P1}^2}{Nb}$

and $Var(\hat{\delta}_2) = \frac{\sigma_{nT}^2}{\xi m_n} + \frac{\sigma_{nP}^2}{(1-\xi)m_n} = \frac{\sigma_{nT}^2}{\xi \hat{r} Nb} + \frac{\sigma_{nP}^2}{(1-\xi)\hat{r}Nb} = \frac{1}{\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)}(\frac{\sigma_{nT}^2}{\xi} + \frac{\sigma_{nP}^2}{(1-\xi)}) = \frac{2}{\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)Nb}(\sigma_{nT}^2 + \sigma_{nP}^2)$ because the probability of being a placebo non-responder in the placebo group at the end of Period 1, $\hat{r}$, is $\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)$ and we have $b = \frac{1}{2}$ for a balanced re-randomization at the beginning of Period 2.

All in all, for an SPD trial, due to zero covariance proved above, the test statistic for H_A against

H_0 is: $Z=\frac{w\hat{\delta}_1+(1-w)\hat{\delta}_2}{\sqrt{w^2Var(\hat{\delta}_1)+(1-w)^2Var(\hat{\delta}_2)}}$.

If N_{NN} is defined as the required sample size for an SPD trial when Period 1 and Period 2 endpoints are normal-normal data, the required sample size should be:

$$\mathrm{N_{NN}}=\frac{(z_{1-\beta}+z_{1-\alpha/2})^2}{(w\delta_1+(1-w)\delta_2)^2/(\,w^2\left(\frac{\sigma_{T1}^2}{(1-b)}+\frac{\sigma_{P1}^2}{b}\right)+(1-w)^2\,\frac{2(\sigma_{nT}^2+\sigma_{nP}^2)}{\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)b})}$$

After collecting data from a trial with an SPD, weighted-z test could be used to assess treatment effect. Lack of data from real trials, a hypothetical trial and its data are used here to illustrate the proposed testing procedure. Assuming there was a phase 2a trial designed to evaluate efficacy, safety and tolerability of experimental drug as an adjunctive treatment for major depressive disorder with significant anxiety symptoms. The weights used for analysis were determined as per the method outlined in Liu et al. (2012) and were 0.846 for Period 1 and 0.154 for Period 2. Based on mixed effect model repeat measurement (MMRM) with treatment(placebo, drug), time and pooled center as factors, time-by-treatment interaction and baseline Hamilton Depression Rating Scale (HDRS17) total score (for respective period) as a covariate, least-square mean differences (SE) in change from baseline to endpoint in HDRS17 from Period 1 and Period 2 for Placebo subjects (Period 1 N=58 and Period 2 N=11) were respectively -9.0 (0.72) and -7.0 (1.62) and for drug group (Period 1 N=61 and Period 2 N=11) were -9.4 (0.72) and -9.8 (1.60) resulting respective Wald test for Period 1 and 2 being -0.5 and -1.2.

$$Z=\frac{w\hat{\delta}_1+(1-w)\hat{\delta}_2}{\sqrt{w^2\mathrm{Var}(\hat{\delta}_1)+(1-w)^2\mathrm{Var}(\hat{\delta}_2)}}=\frac{0.846*(-9.4-(-9.0))+(1-0.846)*(-9.8-(-7.0))}{\sqrt{0.846^2*(0.72^2+0.72^2)+\,(1-0.846)^2*(1.62^2+1.60^2)}}=-0.8274754$$

P-value $\approx$ 0.2. Therefore, based on change from baseline to end point in HDRS17 total score, experimental drug can't be declared to be superior to Placebo as an adjunctive therapy for major depressive disorder with significant anxiety symptoms.

Section 7.2.3 Simulation Results under Assumed Conditional Independence

In the normal-normal case, with 0.8 for power 1- β, 0.6 for weight w and 10 for the mean difference between treatment groups for both periods, the sample size for the SPD is 107 (Table 7.1) while sample size for traditional parallel group design is 126. When mean difference in Period 2 increases to 12, SPD can be more efficient having sample size of only 92, which is a 27% savings relative to parallel group design. Increase of the mean difference from placebo at Period 2 is a reasonable assumption as only placebo non-responders are randomized to Period 2 in SPD. Eliminating placebo responders could possibly increase drug-placebo difference in Period 2. Similar patterns are also observed when w equals to 0.8 or when the power increases to be 0.9.

Although the covariance of $\hat{\delta}_1$ and $\hat{\delta}_2$ is zero in both Chen et al. (2011) and this research, sample size differs little between each. The estimate of probability of being a placebo non-responder in the placebo group at the end of Period 1, $\hat{r}$, is used in Chen et al. (2011) while the expected value of $\hat{r}$ (i.e., $\Phi\left(\frac{c-\mu_{P1}}{\sigma_{P1}}\right)$) is used here. For binomial-binomial data, $E(\hat{r})$= 1-$P_P(r_1)$.

Simulations are done to assess type I error rate and power under the null and alternative hypothesis, respectively, using the sample size calculated in Section 7.2.2. In Column 4 of Table 7.1, the simulated type I error rate and power are displayed next to the sample size N after 10000 runs. For simplicity, type I error rates are simulated under $\mu_{T1} = \mu_{P1} = \mu_{T2} = \mu_{P2} = 15$ for all cases in Table 7.1 while power is simulated under specifications in Columns 3 and 4. Per simulation results, type I error rate has been maintained at one-sided 0.025 level in the presence of simulation error and the designed power of 0.8 (upper half) and 0.9 (lower half) have been achieved in all scenarios. Note that simulations in this section are under assumption of conditional independence because both mean and variance of random variable (Y_{2i}^{nT} and Y_{2i}^{nP} respectively) in Period 2 are not a function of the realization of Period 1 endpoint Y_{1i} , even though both endpoints have occurred on the same set of subjects.

Table 31(Tab. 7.1): Sample size (N) for SPDs

Table 7.1: Sample size (N) for SPDs when Period 1 and Period 2 data are all normally distributed with $X_{1i} \sim \text{Normal}(\mu_{T1},\ \sigma_{T1}^2 = 20^2)$, $X_{1i} \sim \text{Normal}(\mu_{T1},\ \sigma_{T1}^2 = 20^2)$, $Y_{2i}^{nT} | Y_{1i} \leq c \sim \text{Normal}(\mu_{nT},\ \sigma_{nT}^2 = 20^2)$, $Y_{2i}^{nP} | Y_{1i} \leq c \sim \text{Normal}(\mu_{nP},\ \sigma_{nP}^2 = 20^2)$, $\alpha = 0.025$, β =0.1(upper half) or 0.2 (lower half) $c = 7$, $w = 0.6$ or 0.8, the probability of being a placebo non-responder at Period 1 being $E(\hat{r}) = 0.54$, and N_{tpd} denoting corresponding sample size for traditional parallel design.

Power	w	$\delta_1(\mu_{T1},\ \mu_{P1})$	$\delta_2(\mu_{nT},\ \mu_{nP})$	N/simulated type I error rate/power	N_{tpd} with b=0.50	
					$\delta_1(\mu_T,\ \mu_P)$ $1-\beta$	N_{tpd}
$1-\beta = 0.8$	0.6	10 (15, 5)	10 (15, 5)	107/0.0312/0.7906	10 (15, 5) $1-\beta$ =0.8	126
	0.6	10 (15, 5)	12 (15, 3)	92/0.0271/0.7814		
	0.8	10 (15, 5)	10 (15, 5)	104/0.0262/0.7970		
	0.8	10 (15, 5)	12 (15, 5)	96/0.0237/0.7918		
$1-\beta = 0.9$	0.6	10 (15, 5)	10 (15, 5)	143/0.0251/0.8878	10 (15, 5) $1-\beta$ =0.9	169
	0.6	10 (15, 5)	12 (15, 5)	123/0.0250/0.8887		
	0.8	10 (15, 5)	10 (15, 5)	139/0.0284/0.8938		
	0.8	10 (15, 5)	12 (15, 5)	129/0.0274/0.8971		

Section 7.2.4: Simulation Results Under Correlated Endpoints Between Two Periods

Statistical methods illustrated in Section 7.2.3 as well as Chen et al. (2011) and Liu et al. (2012) don't assume dependence structure between endpoints from two periods even though they occur on the same set of subjects. This definitely casts some doubts as in practice we can't rule out dependence when two random variables occur on the same subject. Also, even if the covariance between two phases' estimates is in fact zero, sample covariance may not be zero when the size of the study is small. To address these questions, simulations are conducted for scenarios listed in Table 7.1, while on the contrary conditional dependence is built up accordingly using the properties of the bivariate normal distribution. Given ρ_P being the correlation between Y_{1i} and Y_{2i}^{nP} for subjects who are placebo non-responder in Period 1 and continue to be treated with placebo in Period 2, after observing $\mathrm{Y}_{1\mathrm{i}} = \mathrm{y}_{1\mathrm{i}},$, $\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}$ will be normally distributed with mean $\mu_{\mathrm{nP}} - \rho_P * (\frac{\sigma_{nP}}{\sigma_{P1}}) *(\mathrm{y}_{1\mathrm{i}} - \mu_{\mathrm{p1}})$ and variance σ_{nP}^2*(1-ρ_P^2). Similarly, ρ_{T} and conditional distribution of $\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}$ are defined for subjects who are placebo non-responder in Period 1 and then treated with investigational drug in Period 2. As in Table 7.1, scenarios under null hypothesis are also simulated with $\mu_{\mathrm{T1}} = \mu_{\mathrm{P1}} = \mu_{\mathrm{T2}} = \mu_{\mathrm{P2}} = 15$, but with the conditional mean based on the realized value $\mathrm{y}_{1_\mathrm{i}}$ at the end of Period 1. Using calculated sample size in Table 7.1, type I error rate and power for each scenario are re-simulated using the conditional bivariate normal distribution instead (Table 7.2). Results re-assure maintenance of target power under equal correlations as in Chen et al. (2011), but somehow expose disadvantages of this method under unequal correlations. Simulated power achieves the designed level only for Row 1 with $\rho_{\mathrm{P}} = \rho_{\mathrm{T}} = 0$ and Row 2 with $\rho_{\mathrm{P}} = \rho_{\mathrm{T}} = 0.5$, but lower than designed level in Rows 3-5 when unequal correlation coefficients are $\rho_{\mathrm{P}} = 0.75$ and $\rho_{\mathrm{T}} = 0.5$, $\rho_{\mathrm{P}} = 0.75$ and $\rho_{\mathrm{T}} = 0.25$, and $\rho_{\mathrm{P}} = 0.50$ and $\rho_{\mathrm{T}} = 0.25$, respectively, among which simulated power decreases as the difference between ρ_{P} and ρ_{T} increases. Extensive simulations have been done for other situations but not listed here due to space limitation.

Table 32(Tab. 7.2): Simulated rejection probabilities

Table 7.2: **Simulated rejection probability under null and alternative hypotheses respectively when Period 2 endpoint is conditional upon Period 1 realization.**

ρ_{P}/ρ_T	$w = 0.6$ $1 - \beta = 0.8$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 10(15{,}5)$ N=107 Simulated Type I error rate / power	$w = 0.6$ $1 - \beta = 0.8$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 12(15{,}3)$ N=92 Simulated Type I error rate / power	$w = 0.8$ $1 - \beta = 0.8$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 10(15{,}5)$ N=104 Simulated Type I error rate / power	$w = 0.8$ $1 - \beta = 0.8$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 12(15{,}3)$ N=96 Simulated Type I error rate / power		$w = 0.6$ $1 - \beta = 0.9$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 10(15{,}5)$ N=143 Simulated Type I error rate / power	$w = 0.6$ $1 - \beta = 0.9$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 12(15{,}3)$ N=123 Simulated Type I error rate / power	$w = 0.8$ $1 - \beta = 0.9$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 10(15{,}5)$ N=139 Simulated Type I error rate / power	$w = 0.8$ $1 - \beta = 0.9$ $\delta_1(\mu_{T1}, \mu_{P1})$ $= 10(15{,}5)$ $\delta_2(\mu_{nT}, \mu_{nP})$ $= 12(15{,}3)$ N=129 Simulated Type I error rate / power
0.00 / 0.00	0.0293/0.7905	0.03/0.7893	0.0279/0.7894	0.0275/0.7883		0.0267/0.8877	0.0264/0.893	0.0253/0.8986	0.024/0.8891
0.50 / 0.50	0.0237/0.8147	0.0271/0.814	0.0246/0.7977	0.0271/0.8015		0.0259/0.9111	0.028/0.9114	0.0244/0.9001	0.0229/0.8992

0.75 / 0.50	0.0053/0.719	0.0064/0.731	0.0125/0.7457	0.0131/0.7398		0.0046/0.8279	0.0057/0.8397	0.0106/0.8566	0.0116/0.8632
0.75 / 0.25	0.0001/0.5517	0.0012/0.5762	0.005/0.6703	0.006/0.6669		0.0004/0.6671	0.0005/0.6998	0.004/0.7929	0.0038/0.7946
0.50 / 0.25	0.0077/0.6808	0.0075/0.6966	0.0108/0.7365	0.0127/0.7391		0.0053/0.799	0.0042/0.811	0.0116/0.8523	0.0119/0.8455

Section 7.3: Binomial-Binomial Data

Section 7.3.1: Covariance for $\hat{\delta}_1$ and $\hat{\delta}_2$, Re-examination

X_{1i} ~ Bernoulli$(1, P_T(r_1)$),i=1,…, n; Y_{1i} ~ Bernoulli$(1,\ P_P(r_1)), i = 1, \ldots, m$;

Y_{2i}^{nT}| NR ~ Bernoulli $(1,\ P_{nP}(r_2|nr_1))$, i = 1, …, ξm_n and NR denotes non-responder.

Y_{2i}^{nP}| NR ~ Bernoulli $(1,\ P_{nT}(r_2|nr_1)), i = \xi m_n + 1, \ldots, m_n$, with $P_T(r_1)$ as the probability of being a responder for drug-treated subjects in Period 1, $P_P(r_1)$ as the probability of being a responder for placebo-treated subjects in Period 1, $P_{np}(r_2|nr_1)$ as the probability of being a responder at end of Period 2 when a Period 1 placebo non-responder was treated with placebo in Period 2, and $P_{nT}(r_2|nr_1)$ as the probability of being a responder at end of Period 2 when a Period 1 placebo non-responder was treated with placebo in Period 2.

$\hat{P}_T(r_1)=\frac{1}{n}\sum_{i=1}^{n} X_{1i}$, $\hat{P}_P(r_1)=\frac{1}{m}\sum_{i=1}^{m} Y_{1i}$

$\hat{P}_{nT}(r_2|nr_1) = \frac{1}{\xi m_n}\sum_{i=1}^{\xi m_n} Y_{2i}^{nT}$, $\hat{P}_{nP}(r_2|nr_1) = \frac{1}{(1-\xi)m_n}\sum_{i=\xi m_n+1}^{m_n} Y_{2i}^{nP}$

$cov(\hat{\delta}_1, \hat{\delta}_2) = cov\left(\hat{P}_T(r_1) - \hat{P}_P(r_1), \hat{P}_{nT}(r_2|nr_1) - \hat{P}_{nP}(r_2|nr_1)\right)$

= cov ($\hat{P}_P(r_1)$, $\hat{P}_{nP}(r_2|nr_1)$) - cov($\hat{P}_P(r_1)$, $\hat{P}_{nT}(r_2|nr_1)$)

=0 per proof in Appendix 7.2.

Section 7.3.2 Sample Size Derivation and Evaluation

For binomial data, $Var(\hat{\delta}_1) = \frac{P_T(r_1)(1\text{-}P_T(r_1))}{n} + \frac{P_P(r_1)(1\text{-}P_P(r_1))}{m} = \frac{P_T(r_1)(1\text{-}P_T(r_1))}{N(1-b)} + \frac{P_P(r_1)(1\text{-}P_{TP}(r_1))}{Nb}$.

With the probability of being a non-responder in the placebo group at the end of Period 1 being

1-$P_P(r_1)$, that is E($\hat{r}$)= 1-$P_P(r_1)$, $Var(\hat{\delta}_2)$=$\frac{P_{nT}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nT}(r_2\,|\mathrm{n}r_1))}{\xi m_n} + \frac{P_{nP}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nP}(r_2\,|\mathrm{n}r_1))}{(1-\xi)m_n}$

$$=\frac{P_{nT}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nT}(r_2\,|\mathrm{n}r_1))}{\xi\hat{r}Nb} + \frac{P_{nP}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nP}(r_2\,|\mathrm{n}r_1))}{(1-\xi)\hat{r}Nb}$$

$$= \frac{1}{(1-P_P(r_1))\mathrm{Nb}}\left(\frac{P_{nT}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nT}(r_2\,|\mathrm{n}r_1))}{\xi} + \frac{P_{nP}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nP}(r_2\,|\mathrm{n}r_1))}{(1-\xi)}\right)$$

$$=\frac{2}{(1-P_P(r_1))\mathrm{Nb}}\left(P_{nT}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nT}(r_2\,|\mathrm{n}r_1)) + \; P_{nP}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{np}(r_2\,|\mathrm{n}r_1)\right) \text{, with } b = \frac{1}{2}$$

The sample size for normal-normal data is $\frac{(\mathrm{z}_{1-\beta}+\mathrm{z}_{1-\alpha/2})^2}{\delta^2/\left(\frac{\sigma_{\mathrm{T}}^2}{(1-\mathrm{b})}+\frac{\sigma_{\mathrm{P}}^2}{\mathrm{b}}\right)}$ whereas for binomial-binomial data the sample size is$\left(\mathrm{z}_{1-\beta} + \mathrm{z}_{1-\frac{\alpha}{2}}\right)^2 /[\delta^2/\ \left(\frac{\mathrm{P_T(r_1)(1\text{-}P_T(r_1))}}{(1-\mathrm{b})} + \frac{\mathrm{P_P(r_1)(1\text{-}P_P(r_1))}}{\mathrm{b}}\right)$].

If N_{BB} is defined as the required sample size for an SPD when Period 1 and Period 2 endpoints are binomial-binomial, it should be:

$$N_{BB} = \left(\mathrm{z}_{1-\beta} + \mathrm{z}_{1-\frac{\alpha}{2}}\right)^2 /[(w\delta_1 + (1-w)\delta_2)^2/(\ w^2 \left(\frac{P_T(r_1)(1\text{-}P_T(r_1))}{(1-b)} + \frac{P_P(r_1)(1\text{-}P_P(r_1))}{b}\right) \;+$$

$$(1-w)^2\ \frac{2(\ P_{nT}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nT}(r_2\,|\mathrm{n}r_1)){+}\,P_{nP}(r_2\,|\mathrm{n}r_1)(1\text{-}\,P_{nP}(r_2\,|\mathrm{n}r_1))\ )}{(1-P_P(r_1))\mathrm{b}}\ ]$$

Table 7.3 exhibits sample size for a SPD trial when data are binomially distributed in both periods. Let's take power of 0.8 as an example. Surprisingly, there is no much saving relative to fixed sample design (155 vs. 157) when the rate difference, that is 0.2, is the same in both periods and weight w is 0.6. However, in the case where enrichment is functioning and the rate difference increases from 0.2 in Period 1 to 0.3 in Period 2, the sample size becomes 109, 31% reduction in sample size relative to the corresponding parallel group design. When the rate difference is 0.2 for both periods while weight w is 0.8 with more weight allocated to Period 1 data, sample size decreases to 129. Among four scenarios for power of 0.8, the smallest SPD sample size of 107, is achieved when $\delta_1 = 0.2$, $\delta_2 = 0.3$ and $\mathrm{w} = 0.8$. In summary, different from normal-normal cases, there is almost no sample size saving relative to the traditional parallel group design with $\delta_1 = 0.2,\ \delta_2 = 0.2$ and $w = 0.6$. Simulations under dependence structure are not done because we lack of clear guides on how binomial endpoints from two periods are correlated in practice.

Table 33(Tab. 7.3): Sample sizes

Table 7.3: Sample size when Period 1 and Period 2 data are normally distributed with X_{1i}~Bernoulli($P_T(r_1)$), Y_{1i}~Bernoulli($P_P(r_1)$), Y_{2i}^{nT}| NR~ Bernoulli ($P_{nP}(r_2|nr_1)$),Y_{2i}^{nP}| NR~ Bernoulli ($P_{nT}(r_2|nr_1)$), $\alpha = 0.025, \beta = 0.1\ or\ 0.2,\ w$=0.6 or 0.8, $E(\hat{r}) = 0.4$, and N_{tpd} denoting corresponding sample size for tranditional parallel design.

Power	w	$\delta_1(P_T\ (r_1),$ $P_P\ (r_1))$	$\delta_2(P_{nT}(r_2\ \|nr_1),$ $P_{nP}(r_2\ \|nr_1))$	N /simulated type I error rate/power	N_{tpd} with b=0.50	
					$\delta(P_T\ (r_1),\ P_P\ (r_1))$ $1-\beta$	N_{tpd}
$1-\beta = 0.8$	0.6	0.2 (0.8,0.6)	0.2 (0.8,0.6)	155 /0.1130/0.9363	0.2 (0.8,0.6) $1-\beta$ =0.8	157
	0.6	0.2 (0.8,0.6)	0.3 (0.8, 0.5)	109 /0.1099/0.9323		
	0.8	0.2 (0.8,0.6)	0.2 (0.8, 0.6)	129 /0.0419/0.8342		
	0.8	0.2 (0.8,0.6)	0.3 (0.8, 0.5)	107 /0.0418/0.8295		
$1-\beta = 0.9$	0.6	0.2 (0.8,0.6)	0.2 (0.8, 0.6)	207 /0.1091/0.9756	0.2 (0.8,0.6) $1-\beta$ =0.9	211
	0.6	0.2 (0.8,0.6)	0.3 (0.8, 0.5)	146 /0.1118/0.9698		
	0.8	0.2 (0.8,0.6)	0.2 (0.8, 0.6)	173 /0.0385/0.9207		
	0.8	0.2 (0.8,0.6)	0.3 (0.8, 0.5)	143 /0.0403/0.9145		

Section 7.4: Discussion

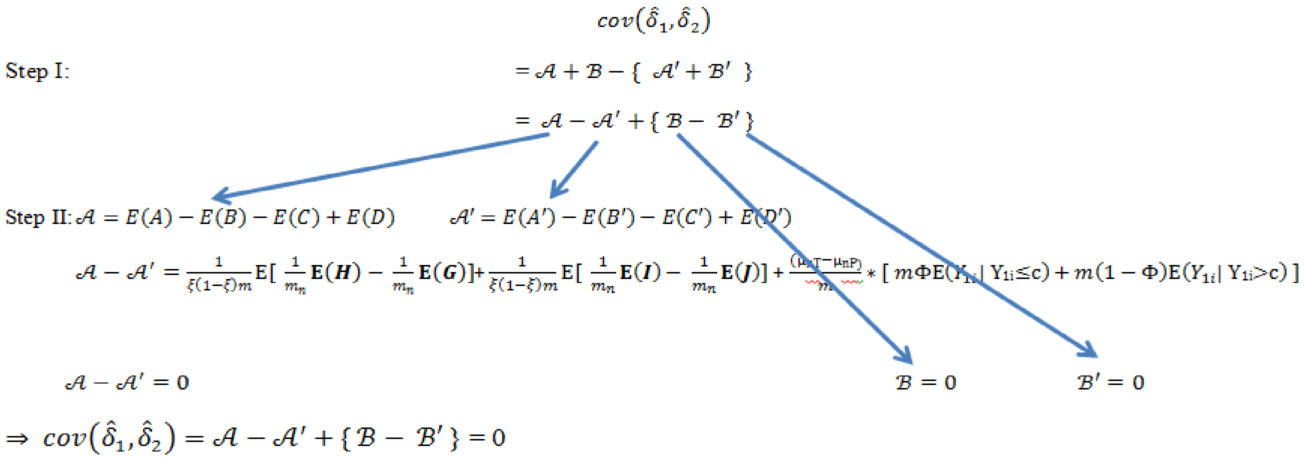
Different from Chen et al. (2011), the covariance of $\hat{\delta}_1$ and $\hat{\delta}_2$ is evaluated to be zero in this paper under rigorous distributional assumptions while without assuming equal correlation coefficients. In derivation, we iteratively used the following formulations: 1) Covariance of two random variables is equal to expectation of conditional covariance plus covariance of conditional expectation. That is, cov(A,B)=E[cov(A|C)]+cov(E(A|C),E(B|C)) where A and B are variables of interest and C is the random variable that A and B to be conditioning upon. 2) Covariance of two random variables is the expectation of the product of expectation of each variable minus its expectation. That is: cov(A, B)=E[(A-E(A|C)) *(B|C-E(B|C))]. Additionally, different from

Chen et al. (2011), the estimated probability of being a placebo non-responder in Period 1 in sample size formula is replaced by its expected value for a better sample size calculation. Zero covariance reduces calculation of variance of the weighted estimator from three components to two components and the power of proposed method is confirmed in Table 7.1 under the conditional independence assumption. However, further simulations in Table 7.2 under conditional dependence show the limitation of this proposed method but point out the direction of future research. Rigorous formulation is in need for correlated endpoints from the two periods in a SPD trial. Besides normal-normal data, binomial-binomial data have also been explored in Section 7.3. Substantial saving of sample size, more than 30%, is achieved in normal-normal data but not in binomial-binomial data. We also observed that further savings is achieved when the weight increased from 0.6 to 0.8 and more weights is placed on Period 1 for normal-normal data. Impacts from weight change/weight optimization and normal-binomial data and binomial-normal data have also been investigated by authors but not shown due to space limitation. All in all, this paper provides another view of combination test in an SPD trial and rigorously formulates covariance calculation without equal correlation coefficients. Most importantly it investigates the performance of the proposed method under unequal correlation coefficients in addition to independence assumption, which haven’t been done by either Chen et al. (2011) or Liu et al. (2012).

Reference

Chen YF, Yang Y, Hung HMJ, Wang SJ. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemp. Clin. Trials* 2011; 32: 592-604.
Fava M, Evins AE, Dorer DJ, Schoenfeld DA. The problem of the placebo response in clinical trials for psychinatric disorders: culprits, possible remedies, and a novel study design approach. *Psychother Psychosom* 2003; 72:115-27.
-----(2004), “Errarum”, *Psychotherapy and Psychosomatics,* 73, 123.
Liu Q, Lim P, Singh J, Lewin D, Schwab B and Kent J. Doubly randomization delayed start design for enrichment studies with responders or non-responders. *Journal of Biopharmaceutical Statistics* 2012; 22:4, 737-757.
Tamura R, Huang X. An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clin Trials* 2007; 309-17.
Trivedi MH, Rush AI. Does a placebo run-in or a placebo treatment cell affect the efficiency of antidepressant medications? *Neuro psycho pharmacology* 1994; 11: 33-43.
Walsh BT, Seidman SN, Sysko R, Could M. Placebo response in studies of major depression: variable, substantial and growing. *JAMA* 2002; 287:1840-7
Zellner A. An efficient method of estimating seeming unrelated regressions and tests for aggregation bias. *J Am Stat Assoc* 1962; 57:348-68.

Appendix 7.1 : covariance for Normal-Normal Case



$cov(\hat{\delta}_1,\hat{\delta}_2)$

Step I: $= \mathcal{A} + \mathcal{B} - \{ \mathcal{A}' + \mathcal{B}' \}$

$= \mathcal{A} - \mathcal{A}' + \{ \mathcal{B} - \mathcal{B}' \}$

Step II: $\mathcal{A} = E(A) - E(B) - E(C) + E(D)$ $\quad \mathcal{A}' = E(A') - E(B') - E(C') + E(D')$

$\mathcal{A} - \mathcal{A}' = \frac{1}{\xi(1-\xi)m} \mathrm{E}[\, \frac{1}{m_n}\mathbf{E}(\boldsymbol{H}) - \frac{1}{m_n}\mathbf{E}(\boldsymbol{G})] + \frac{1}{\xi(1-\xi)m}\mathrm{E}[\, \frac{1}{m_n}\mathbf{E}(\boldsymbol{I}) - \frac{1}{m_n}\mathbf{E}(\boldsymbol{J})] + \frac{(\mu_{nT}-\mu_{nP})}{m} * [\, m\Phi \mathrm{E}(Y_{1i}|\, Y_{1i} \leq c) + m(1-\Phi)\mathrm{E}(Y_{1i}|\, Y_{1i} > c)\,]$

$\mathcal{A} - \mathcal{A}' = 0$ $\qquad \mathcal{B} = 0$ $\qquad \mathcal{B}' = 0$

$\Rightarrow cov(\hat{\delta}_1,\hat{\delta}_2) = \mathcal{A} - \mathcal{A}' + \{ \mathcal{B} - \mathcal{B}' \} = 0$

Illustration 7.1: A schematic of the proof of zero covariance in normal-normal case.

$cov(\hat{\delta}_1,\hat{\delta}_2) = cov(\hat{\mu}_{T1} - \hat{\mu}_{P1}, \hat{\mu}_{nT} - \hat{\mu}_{nP})$
$= cov(\hat{\mu}_{T1},\hat{\mu}_{nT}) - cov(\hat{\mu}_{T1},\hat{\mu}_{nP}) - cov(\hat{\mu}_{P1},\hat{\mu}_{nT}) + cov(\hat{\mu}_{P1},\hat{\mu}_{nP})$
$= cov(\hat{\mu}_{P1},\hat{\mu}_{nP}) - cov(\hat{\mu}_{P1},\hat{\mu}_{nT})$, with $cov(\hat{\mu}_{T1},\hat{\mu}_{nT})$ and $cov(\hat{\mu}_{T1},\hat{\mu}_{nP})$ being zero as they are on different subjects
$= E[cov(\hat{\mu}_{P1},\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m)] +$
$cov\big(E(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m), E(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m)\big) -$
$\{ \; E[cov(\hat{\mu}_{P1},\hat{\mu}_{nT}|I(Y_{1i}>c),\; i = 1,\ldots,m)] +$
$cov\big(E(\hat{\mu}_{P1}|I(Y_{1i}>c),\; i = 1,\ldots,m),\; E(\hat{\mu}_{nT}|I(Y_{1i}>c),\; i = 1,\ldots,m)\big) \; \}$
$= \mathcal{A} + \mathcal{B} - \{ \mathcal{A}' + \mathcal{B}' \}$
Let $\mathcal{A} = E[cov(\hat{\mu}_{P1},\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m)]$, then the inner part of this expectation is as follows:
$cov(\hat{\mu}_{P1},\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m)$
$= cov\big((\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m), (\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \big)$
$= \mathrm{E}[\; ((\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m) - \mathrm{E}(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)) *$
$(\; (\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) - \mathrm{E}(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m)) \;\;]$
$= \mathrm{E}\,[\;\; (\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) -$
$(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)\mathrm{E}(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \; -$
$\mathrm{E}(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) +$
$\mathrm{E}(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)\mathrm{E}(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;\;]$
$= \mathrm{E}[\; (\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;] -$
$\mathrm{E}[\; (\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)\mathrm{E}(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;] -$
$\mathrm{E}[\; \mathrm{E}(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;] +$
$\mathrm{E}[\; \mathrm{E}(\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)\mathrm{E}(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;]$
$= A - B - C + D$
So $\mathcal{A} = E(A) - E(B) - E(C) + E(D)$
A $= E[\; (\hat{\mu}_{P1}|I(Y_{1i}>c), i = 1,\ldots,m)(\hat{\mu}_{nP}|I(Y_{1i}>c), i = 1,\ldots,m) \;]$

$= \mathrm{E}[\ \frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c}+\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})(\frac{1}{(1-\xi)m_n}(\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}))\ \ ]$

$=$

$\mathrm{E}[\frac{1}{(1-\xi)mm_n}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c}^*\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}+\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c}^*\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}\ )]$

$\mathrm{B}\ =\ \mathrm{E}[\ (\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\mathrm{E}(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ \ ]$

$\ \ =\ \mathrm{E}[\frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c}+\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})^*\mu_{\mathrm{nP}}\ ]$

$\ \ =\ \frac{\mu_{\mathrm{nP}}}{m}\left(m_nE(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+(m-m_n)E(Y_{1i}|\ \mathrm{Y_{1i}}>c)\right)$

Based on property of truncated normal distribution,

$E(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})=\mu_{\mathrm{P1}}\text{ - }\sigma_{\mathrm{P1}}\frac{\emptyset(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})}{\Phi(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})}$, $E(Y_{1i}|\ \mathrm{Y_{1i}}>c)=\mu_{\mathrm{P1}}+\sigma_{\mathrm{P1}}\frac{\emptyset(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})}{1-\Phi(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})}$, with $\emptyset$ as the standard normal density and Φ as the CDF of standard normal. $1-\Phi\left(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}}\right)=\mathrm{P_P(r_1)}$, probability of being a placebo repsonder at the end of Period 1. For simplicity, let's use $\emptyset$ denote $\emptyset(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})$ and Φ denote $\Phi(\frac{\mathrm{c}-\mu_{\mathrm{P1}}}{\sigma_{\mathrm{P1}}})$ in all subsequent equations instead.

$\mathrm{C}\ =\ \mathrm{E}[\ E(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ ]$

$=\ \mathrm{E}[\ \frac{1}{m}\left(m_nE(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+(m-m_n)E(Y_{1i}|\ \mathrm{Y_{1i}}>c)\right)^*\ (\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ \ ]$

$=\ \frac{1}{m}\mathrm{E}\left(m_nE(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+(m-m_n)E(Y_{1i}|\ \mathrm{Y_{1i}}>c)\right)\mathrm{E}[(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})]$

$=\frac{\mu_{\mathrm{nP}}}{m}\left(m_nE(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+(m-m_n)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>c)\right)$

$\mathrm{D}\ =\ \mathrm{E}[\ \mathrm{E}(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\mathrm{E}(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ ]$

$=\frac{\mu_{\mathrm{nP}}}{m}\left(m_nE(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+(m-m_n)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>c)\right)$

$\therefore\ \mathcal{A}\ =\ E[cov(\hat{\mu}_{\mathrm{P1}},\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})]$

$\qquad=\ E(A)-E(B)-E(C)+E(D)$

$\qquad=\ \frac{1}{(1-\xi)m_n}\mathrm{E}\ [\ \frac{1}{m}(\ E[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c}^*\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}]\ +$

$\qquad\qquad E[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c}^*\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}\ ]\ \ )\ \ ]$

$\qquad-\frac{\mu_{\mathrm{nP}}}{m}*[\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})\ ]$, with Φ defined as above.

Similarly, $\mathcal{A}'\ =\ \mathrm{E}[cov(\hat{\mu}_{\mathrm{P1}},\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})]$

$cov(\hat{\mu}_{\mathrm{P1}},\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})$

$=\mathrm{E}[(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})(\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ ]-$

$\qquad\mathrm{E}[\ \ (\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\mathrm{E}(\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ \ ]-$

$\qquad\mathrm{E}[\ E(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})(\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ ]+$

$\qquad\mathrm{E}[\ E(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\mathrm{E}(\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y_{1i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\ ]$

$=\ A'-B'-C'+D'$

$\mathrm{A'}\ =\ \ \mathrm{E}[\frac{1}{\xi mm_n}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c}^*\sum_{i=1}^{\xi m_n} \mathrm{Y_{2i}^{nT}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}+\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c}^*\sum_{i=1}^{\xi m_n} \mathrm{Y_{2i}^{nT}}|\ \mathrm{Y_{1i}}\leq\mathrm{c}\ )]$

$\mathrm{E(B')}\ =\ \frac{\mu_{\mathrm{nT}}}{m}*[\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})\ ]$

$E(C')\ =\ \frac{\mu_{\mathrm{nT}}}{m}*[\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})\ ]$

$E(D')\ =\ \frac{\mu_{\mathrm{nT}}}{m}*[\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y_{1i}}>\mathrm{c})\ ]$

$\therefore\ \mathcal{A}-\mathcal{A}'\ =\ \ \frac{1}{(1-\xi)m}$

E

$[\frac{1}{m_n}(\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}]\ +$

$\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ )\ ]\ \ -$

$-\frac{1}{\xi m}\mathrm{E}\ [\frac{1}{m_n}(\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}] + \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ )\ ]$

$+\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP}})}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}) + m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c})\ ]$

$=\ \frac{1}{\xi(1-\xi)m}\ \mathrm{E}\ \ \ [\frac{1}{m_n}(\ \xi\mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}] -$

$(1\text{-}\xi)\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}])\ ]\ +$

$\frac{1}{\xi(1-\xi)m}\ \mathrm{E}\ \ \ [\frac{1}{m_n}(\ \xi\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ] -$

$(1\text{-}\xi)\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ )\ \ \ ]+$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP}})}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}) + m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c})\ ]$

$=$

$\frac{1}{\xi(1-\xi)m}\mathrm{E}\ [\frac{1}{m_n}(\ \ \mathrm{E}\ (\xi\ \ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ {*}\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ \ +$

$\xi\ \ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ {*}\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ )\ ) - \frac{1}{m_n}\mathrm{E}\ [\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ {*}\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ \ ]+$

$\frac{1}{\xi(1-\xi)m}\ E\ [\frac{1}{m_n}\ (\ \mathrm{E}\ (\ \xi\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}\ {*}\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ +$

$\xi\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}\ {*}\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ )\ )\ \ -$

$\frac{1}{m_n}\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}\ {*}\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ ]+$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP}})}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}) + m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c})\ ]$

$=\ \frac{1}{\xi(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(H) - \frac{1}{m_n}\mathrm{E}(G)]+\frac{1}{\xi(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(I) - \frac{1}{m_n}\mathrm{E}(J)]\ +$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP}})}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}) + m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c})\ ]$

$\mathrm{E(H)}\ =\ \xi\ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}(\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ +\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c})\ ]$

$=\ \xi\ \mathrm{E}[\ \ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ +\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c})|\ \ \ \ \mathrm{Y}_{1i}|\mathrm{Y}_{1i}\leq \mathrm{c}\ ]\ \ ]$

$=\ \xi\ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\mathrm{E}(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ +\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ )\ ]$

$=\ \xi\ \mathrm{E}[\ (\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}){*}(\ (1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})\ ]$

$=\ \xi(\ (1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})\ m_n\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c})$

$\mathrm{E(G)}\ =\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}]$

$=\ \mathrm{E}[\ \mathrm{E}\ [\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}|\ \mathrm{Y}_{1i}|\mathrm{Y}_{1i}\leq \mathrm{c}\ \ ]\ ]$

$=\mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c}{*}\ \mathrm{E}(\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ \ )\ \ \ ]$

$=\ \xi m_n\mu_{\mathrm{nT}}m_n\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1i}\leq \mathrm{c})$

$\mathrm{E(I)}\ =\ \xi\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}(\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c} + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ )]$

$=\ \xi\mathrm{E}[\ \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c} + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ )|\mathrm{Y}_{1i}|\mathrm{Y}_{1i}\leq \mathrm{c}]]$

$=\ \xi\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}\mathrm{E}[(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2i}^{\mathrm{nP}}|\ \mathrm{Y}_{1i}\leq \mathrm{c} + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2i}^{\mathrm{nT}}|\ \mathrm{Y}_{1i}\leq \mathrm{c}\ )|\mathrm{Y}_{1i}|\mathrm{Y}_{1i}\leq \mathrm{c}]]$

$=\ \xi\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1i}>\mathrm{c}{*}(\ (1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})]$

$= \xi((1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})(m-m_n)\ \mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})$

$\mathrm{E(J)} = \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}^{*}\sum_{i=1}^{\xi m_n}\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c}\ \ ]$

$= \mathrm{E}[\ \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}^{*}\sum_{i=1}^{\xi m_n}\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c}|\mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c}]\ ]$

$= \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c} * \mathrm{E}(\sum_{i=1}^{\xi m_n}\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})\ \ ]$

$= \mathrm{E}\ [(\sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ \ \xi m_n\mu_{\mathrm{nT}}]$

$= \xi m_n\mu_{\mathrm{nT}}(m-m_n)\ \mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})$

$\therefore \mathcal{A}-\mathcal{A}' = \frac{1}{\xi(1-\xi)m}\mathrm{E}\left[\ \frac{1}{m_n}\mathrm{E}(H)-\ \frac{1}{m_n}\mathrm{E}(G)\right]+\frac{1}{\xi(1-\xi)m}\mathrm{E}\left[\ \frac{1}{m_n}\mathrm{E}(I)-\ \frac{1}{m_n}\mathrm{E}(J)\right]+$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP})}}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ ]$

$= \frac{1}{\xi(1-\xi)m}\mathrm{E}\ [\ \frac{1}{m_n}\ \xi((1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})m_n\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})-$

$\frac{1}{m_n}\xi m_n\mu_{\mathrm{nT}}m_n\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})\ \ ]\ +$

$\frac{1}{\xi(1-\xi)m}\mathrm{E}\ [\ \frac{1}{m_n}\xi((1\text{-}\xi)m_n\mu_{\mathrm{nP}}+\xi m_n\mu_{\mathrm{nT}})(m-m_n)\ \mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})-$

$\frac{1}{m_n}\xi m_n\mu_{\mathrm{nT}}(m-m_n)\ \mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ \ ]\ +$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP})}}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ ]$

$= \frac{1}{(1-\xi)m}\ \mathrm{E}\ (Y_{1i}|\mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})\mathrm{E}[(\ (1\text{-}\xi)\mu_{\mathrm{nP}}+\ \xi\mu_{\mathrm{nT}}-\mu_{\mathrm{nT}})m_n]-$

$\frac{1}{(1-\xi)m}\ \mathrm{E}\ (Y_{1i}|\mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\mathrm{E}\left[\left(\ (1\text{-}\xi)\mu_{\mathrm{nP}}+\ \xi\mu_{\mathrm{nT}}-\mu_{\mathrm{nT}}\right)(m-m_n)\right]+$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP})}}{m} * [\ m\Phi\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+m(1-\Phi)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ ]$

$= \mathrm{E}\ (Y_{1i}|\mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})(\ \mu_{\mathrm{nP}}-\mu_{\mathrm{nT}})\ (\ 1-P_p(r_1)\ )+\mathrm{E}(Y_{1i}|\mathrm{Y}_{1\mathrm{i}}>\mathrm{c})(\ \mu_{\mathrm{nP}}-\mu_{\mathrm{nT}})\ P_p(r_1)\ +$

$\frac{(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP})}}{m} * \left[\ m\left(1-P_p(r_1)\right)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+mP_p(r_1)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ \right]$

$= \mathrm{E}\ (Y_{1i}|\mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})(\ \mu_{\mathrm{nP}}-\mu_{\mathrm{nT}})(1-\ P_p(r_1))+\mathrm{E}(Y_{1i}|\mathrm{Y}_{1\mathrm{i}}>\mathrm{c})(\ \mu_{\mathrm{nP}}-\mu_{\mathrm{nT}})P_p(r_1)\ +$

$(\mu_{\mathrm{nT}}-\mu_{\mathrm{nP}}) * [\ (1-\ P_p(r_1))\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+P_p(r_1)\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c})\ ] = 0$

$\mathcal{B} = cov\left(\ \mathrm{E}(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m}),\mathrm{E}(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\right)$

$= \mathrm{E}\left[\ \left(\mathrm{E}\ (\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})-\ \mathrm{E}(\hat{\mu}_{\mathrm{P1}})\right)\left(\mathrm{E}\ (\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})-\ \mathrm{E}(\hat{\mu}_{\mathrm{nP}})\right)\ \right]$

$= \mathrm{E}\ [\ (\frac{1}{m}(\ \sum_{i=1}^{m_n}E(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+\sum_{i=m_n+1}^{m}\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}))\text{-}\ \mu_{\mathrm{P1}}\ )\ \ (\frac{1}{(1-\xi)m_n}\ \sum_{i=\xi m_n+1}^{m_n}\mathrm{E}(\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})$

$-\mathrm{E}(\hat{\mu}_{\mathrm{nP}})\quad)\ \ ] = \mathrm{E}\ [\ (\frac{1}{m}(\ \sum_{i=1}^{m_n}\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+\sum_{i=m_n+1}^{m}\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}))\text{-}\ \mu_{\mathrm{P1}}\ )\ \ (\mu_{\mathrm{np}}-\mu_{\mathrm{np}}\ )\ \ ]=0$

$\mathcal{B}' = cov\left(\ E(\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m}),\mathrm{E}(\hat{\mu}_{\mathrm{nP}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})\right)$

$= \mathrm{E}\left[\ \left(E\ (\hat{\mu}_{\mathrm{P1}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})-\ \mathrm{E}(\hat{\mu}_{\mathrm{P1}})\right)\left(E\ (\hat{\mu}_{\mathrm{nT}}|I(\mathrm{Y}_{1\mathrm{i}}>\mathrm{c}),\mathrm{i}=1,\dots,\mathrm{m})-\ \mathrm{E}(\hat{\mu}_{\mathrm{nT}})\right)\ \right]$

$= \mathrm{E}\ [\ (\frac{1}{m}(\ \sum_{i=1}^{m_n}E(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+\sum_{i=m_n+1}^{m}\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}))\text{-}\ \mu_{\mathrm{P1}}\ )\ \ (\frac{1}{(1-\xi)m_n}\ \sum_{i=\xi m_n+1}^{m_n}\mathrm{E}(\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})$

$-\mathrm{E}(\hat{\mu}_{\mathrm{nT}})\quad)\ \ ] = \mathrm{E}\ [\ (\frac{1}{m}(\ \sum_{i=1}^{m_n}E(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}\leq\mathrm{c})+\sum_{i=m_n+1}^{m}\mathrm{E}(Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}>\mathrm{c}))\text{-}\ \mu_{\mathrm{P1}}\ )\ \ (\mu_{\mathrm{nT}}-\mu_{\mathrm{nT}}\ )\ \ ]=0$

Thus $cov\left(\hat{\delta}_1,\hat{\delta}_2\right) = \mathcal{A}+\mathcal{B}-(\mathcal{A}'+\mathcal{B}') = \mathcal{A}-\mathcal{A}'=0$ for Normal-Normal scenario.

Appendix 7.2: covariance for Binomial-Binomial Case

$cov\left(\hat{\delta}_1, \hat{\delta}_2\right) = cov\left(\hat{P}_T(r_1) - \hat{P}_P(r_1), \hat{P}_{nT}(r_2|nr_1) - \hat{P}_{nP}(r_2|nr_1)\right)$

$= cov(\ \hat{P}_P(r_1),\ \hat{P}_{nP}(r_2|nr_1))\ - cov(\ \hat{P}_P(r_1),\ \hat{P}_{nT}(r_2|nr_1))$

$= \mathrm{E}\left[cov(\ \hat{P}_P(r_1), \hat{P}_{nP}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right] +$

$cov\left(\ \mathrm{E}(\ \hat{P}_P(r_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}), \mathrm{E}(\hat{P}_{nP}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right) -$

$\{\ \ \mathrm{E}\left[cov(\ \hat{P}_P(r_1), \hat{P}_{nT}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right] +$

$cov\left(\ \mathrm{E}(\ \hat{P}_P(r_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}), \mathrm{E}(\hat{P}_{nT}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right)\ \ \}$

$= \mathcal{A} + \mathcal{B} - (\mathcal{A}' + \mathcal{B}')$

$\mathcal{A} = \mathrm{E}\left[cov(\ \hat{P}_P(r_1), \hat{P}_{nP}(r_2|nr_1))|\ I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right]$, the inner part of the expectation is as follows:

$cov(\ \hat{P}_P(r_1), \hat{P}_{nP}(r_2|nr_1)|\ I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})$

$= cov(\ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}), (\hat{P}_{nP}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}))$

$= \mathrm{E}\ [\ \ (\ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}) - \mathrm{E}(\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ ) *$

$(\ \ (\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}) - \mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ )\ \ \ ]$

$= \mathrm{E}\ [\ \ \ \ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}) -$

$(\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ -$

$\mathrm{E}\ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}) +$

$\mathrm{E}\ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ \ ]$

$= \mathrm{E}\left[\ \ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ \right]-$

$\mathrm{E}\ [\ \ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ ] -$

$\mathrm{E}\ [\ \ \mathrm{E}(\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ ] +$

$\mathrm{E}\ [\ \ \mathrm{E}(\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ ]$

$= \ A - B - C + D$

$\mathcal{A}\ =\ E(A) - E(B) - E(C) + E(D)$

$\mathrm{A}\ =\ \mathrm{E}\left[\ \ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ \ \right]$

$=\ \mathrm{E}\ [\ \ \frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}=0 + \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}=1)(\frac{1}{(1-\xi)m_n}(\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}=0))\ \ \ ]$

$=\ \ \mathrm{E}$

$[\frac{1}{(1-\xi)mm_n}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}=0 * \sum_{i=\xi m_n}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}=0 + \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}=1 * \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y_{2i}^{nP}}|\ \mathrm{Y_{1i}}=0\ )]$

$\mathrm{B}\ =\ \mathrm{E}\ [\ \ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=0), \mathrm{i} = 1, \dots, \mathrm{m})\mathrm{E}(\hat{P}_{nP}(r_2|nr_1)\big|I(\mathrm{Y_{1i}}=0), \mathrm{i} = 1, \dots, \mathrm{m})\ \ ]$

$= \mathrm{E}\ [\frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y_{1i}}=0 + \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y_{1i}}=1) * P_{nP}(r_2|nr_1)\ ]$

$= \ \frac{P_{nP}(r_2|nr_1)}{m} E[m_n * 0 + (m - m_n) * 1]$

$= \ P_{nP}(r_2|nr_1)\ P_P(r_1)$

$\mathrm{C} = \mathrm{D} = \ P_{np}(r_2|nr_1)\ P_P(r_1)$

Similarly,

$\mathcal{A}' = \mathrm{E}\left[cov(\ \hat{P}_P(r_1), \hat{P}_{nT}(r_2|nr_1))|\ I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\right]$

$cov(\ \hat{P}_P(r_1), \hat{P}_{nT}(r_2|nr_1)|\ I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})$

$= cov(\ (\ \hat{P}_P(r_1)\big|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}), (\hat{P}_{nT}(r_2|nr_1)|I(\mathrm{Y_{1i}}=1), \mathrm{i} = 1, \dots, \mathrm{m})\ )$

$= \mathrm{E}\ [\ \big(\big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big) - \mathrm{E}\big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ \big) *$

$(\ \big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big) - \mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ )\quad]$

$= \mathrm{E}\ [\quad \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big) -$

$\big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ -$

$\mathrm{E}\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big) +$

$\mathrm{E}\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\quad]$

$= \mathrm{E}\ \big[\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ \big]-$

$\mathrm{E}\ [\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ ] -$

$\mathrm{E}\ [\ E\big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ ] +$

$\mathrm{E}\ [\ E\big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ ]$

$= \ A' - B' - C' + D'$

$A\text{'} = \ E\ \big[\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}=1}), \mathrm{i} = 1, \dots, \mathrm{m}\big)\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}=}\mathrm{c}), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ \big]$

$= \mathrm{E}\ [\ \frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0+ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1)(\frac{1}{(1-\xi)m_n}(\sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0))\quad]$

$= \ \mathrm{E}$

$[\frac{1}{(1-\xi)mm_n}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0^* \sum_{i=\xi m_n}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}\ =0+ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1^* \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ )]$

$\mathrm{B}\text{'}\ = \ \mathrm{E}\ [\ \big(\hat{P}_P(r_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=0), \mathrm{i} = 1, \dots, \mathrm{m}\big)\mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1)\big|I(\mathrm{Y}_{1\mathrm{i}}=0), \mathrm{i} = 1, \dots, \mathrm{m}\big)\ ]$

$= \ \mathrm{E}\ [\frac{1}{m}(\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0+ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1)^* P_{nT}(r_2|nr_1)\ ]$

$= \ \frac{P_{nT}(r_2|nr_1)}{m} E[m_n * 0 + (m - m_n) * 1]$

$= \ (r_2|nr_1)\ P_P(r_1)$

$\mathrm{C}\text{'} = \mathrm{D}\text{'} = \ P_{nT}(r_2|nr_1)\ P_P(r_1)$

$\therefore \mathcal{A} - \mathcal{A}' = \ \frac{1}{(1-\xi)m}$

E

$[\frac{1}{m_n}(\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0^* \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0]\ +$

$\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1^* \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ ]\ )\ ]\quad -$

$-\frac{1}{\xi m}\mathrm{E}\ [\frac{1}{m_n}(\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0^* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0] + \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1^* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ ]\ )\ ]$

$+(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$= \ \frac{1}{\xi(1-\xi)m}\ \mathrm{E}\quad [\frac{1}{m_n}(\ \xi\mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0^* \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0]\ \ -$

$(1\text{-}\xi)\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0^* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0])\ ]\ +$

$\frac{1}{\xi(1-\xi)m}\ \mathrm{E}\quad [\frac{1}{m_n}(\ \xi\mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1^* \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ ] -$

$(1\text{-}\xi)\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1^* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ ]\ )\quad]\ +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$=$

$\frac{1}{\xi(1-\xi)m}\mathrm{E}\ [\ \frac{1}{m_n}(\quad \mathrm{E}\ (\xi\ \ \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0\ *\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ \ +$

$\xi \sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0\ *\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ )\ ) - \ \frac{1}{m_n}\mathrm{E}\ [\sum_{i=1}^{m_n} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=0\ * \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ ]\ \ ] +$

$\frac{1}{\xi(1-\xi)m}\ \ E\ [\ \frac{1}{m_n}\ (\ \ \mathrm{E}\ (\ \xi \sum_{i=m_n+1}^{m} Y_{1i}|\ \mathrm{Y}_{1\mathrm{i}}=1\ *\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}}|\ \mathrm{Y}_{1\mathrm{i}}=0\ +$

$\xi\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1\ *\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ ]\ )\ ) -$

$\frac{1}{m_n}\ \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1\ *\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ ]\ ] +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$= \ \frac{1}{\xi(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(H) - \ \frac{1}{m_n}\mathrm{E}(G)] + \frac{1}{\xi(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(I) - \ \frac{1}{m_n}\mathrm{E}(J)]\ +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$\mathrm{E(H)}\ =\ \xi\ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0*(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0)\ ]$

$=\ \xi\ \mathrm{E}[\ \ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0*(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0) \mid \ \ \mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}=0\ ]\ \ ]$

$=\ \xi\ \mathrm{E}[\ \sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0*\mathrm{E}(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ )\ ]$

$= \xi\ \mathrm{E}[\ (\sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0)*(\ (1\text{-}\xi)m_n P_{nP}(r_2|nr_1) + \xi m_n P_{nT}(r_2|nr_1))\ ]$

$=\ \xi(\ (1\text{-}\xi)m_n P_{nP}(r_2|nr_1) + \xi m_n P_{nT}(r_2|nr_1))\ m_n * 0 = 0$

$\mathrm{E(G)}\ =\ \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0]$

$=\ \mathrm{E}[\ \mathrm{E}\ [\sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0 \mid \mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}=0\ \ ]\ ]$

$= \mathrm{E}[\sum_{i=1}^{m_n} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=0*\ \mathrm{E}(\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ \ )\ \ \ ]$

$= \xi m_n P_{nT}(r_2|nr_1) m_n * 0 = 0$

$\mathrm{E(I)}\ =\ \xi \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1*(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0 + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ )]$

$=\ \xi \mathrm{E}[\ \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1*(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0 + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ )|\mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}=0]]$

$=\ \xi \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1*\mathrm{E}[(\ \sum_{i=\xi m_n+1}^{m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} \mid \mathrm{Y}_{1\mathrm{i}}=0 + \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ )|\mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}=0]]$

$=\ \xi \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1*(\ (1\text{-}\xi)m_n P_{nP}(r_2|nr_1) + \xi m_n P_{nT}(r_2|nr_1))]$

$=\ \xi((1\text{-}\xi)m_n P_{nP}(r_2|nr_1) + \xi m_n P_{nT}(r_2|nr_1))(m - m_n)$

$\mathrm{E(J)}\ = \mathrm{E}[\ \sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ \ \ ]$

$=\ \mathrm{E}[\ \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1* \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0|\mathrm{Y}_{1\mathrm{i}}|\mathrm{Y}_{1\mathrm{i}}=0]\ ]$

$=\ \mathrm{E}[\sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1 * \mathrm{E}(\ \sum_{i=1}^{\xi m_n} \mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} \mid \mathrm{Y}_{1\mathrm{i}}=0\ )\ \ ]$

$=\ \mathrm{E}[(\sum_{i=m_n+1}^{m} Y_{1i} \mid \mathrm{Y}_{1\mathrm{i}}=1)\ \ \xi P_{nT}(r_2|nr_1)]$

$= \xi m_n P_{nT}(r_2|nr_1)(m - m_n)$

$\therefore \mathcal{A} - \mathcal{A}' = \ \ \frac{1}{\xi(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(H) - \ \frac{1}{m_n}\mathrm{E}(G)] + \frac{1}{a(1-\xi)m}\mathrm{E}[\ \frac{1}{m_n}\mathrm{E}(I) - \ \frac{1}{m_n}\mathrm{E}(J)]\ +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$=\ \frac{1}{\xi(1-\xi)m}\ \mathrm{E}\left[\ \frac{1}{m_n}\mathrm{E}(I) - \ \frac{1}{m_n}\mathrm{E}(J)\right]\ +\ (P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$=$

$\frac{1}{\xi(1-\xi)m}\ \mathrm{E}\left[\frac{1}{m_n}\ \ \xi\ ((1\text{-}\xi)m_n P_{nP}(r_2/nr_1) + \xi m_n P_{nT}(r_2/nr_1))\ (m - m_n)\ -\right.$

$\left.\frac{1}{m_n}\ (\ \xi m_n P_{nT}(r_2/nr_1)(m - m_n)\ \ )\right] +\ (P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$=\ \frac{1}{(1-\xi)m}\ \mathrm{E}[\ \ ((1\text{-}\xi)P_{nP}(r_2|nr_1) + \xi P_{nT}(r_2|nr_1))\ (m - m_n)\ \ -\ \ (\ P_{nT}(r_2|nr_1)\ (m - m_n)\ \ )] +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$= \frac{1}{(1-\xi)m}\left[\ \left((1\text{-}\xi)P_{nP}(r_2|nr_1) + aP_{nT}(r_2|nr_1)\ \right) -\ P_{nT}(r_2|nr_1)\ \right]\ \mathrm{m}P_P(r_1)\ +$

$(P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1)$

$=\ \text{-}\ (P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1) + (P_{nT}(r_2|nr_1) - P_{nP}(r_2|nr_1))P_P(r_1) = 0$

$\mathcal{B} = cov\Big(\mathrm{E}\big(\hat{P}_P(r_1) | I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\big), \mathrm{E}\big(\hat{P}_{nP}(r_2|nr_1) | I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\Big)$

$= \mathrm{E}\Big[\Big(\mathrm{E}\left(\hat{P}_P(r_1) \middle| I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\right)$

$\qquad\qquad - \mathrm{E}\big(\hat{P}_P(r_1)\big)\Big), \Big(\mathrm{E}\left(\hat{P}_{nP}(r_2|nr_1) \middle| I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\right) - \mathrm{E}\Big(\hat{P}_{nP}(r_2|nr_1)\Big)\Big) \Big]$

$= \mathrm{E}\ [\ (\frac{1}{m} (\sum_{i=1}^{m_n} Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}0 + \sum_{i=m_n+1}^{m} Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}1) \text{-}\ P_P(r_1)\ )\ \ (\frac{1}{(1-\xi)m_n} \sum_{i=\xi m_n+1}^{m_n} \mathrm{E}(\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nP}} |\ \mathrm{Y}_{1\mathrm{i}}{=}0) -$

$\mathrm{E}\Big(\hat{P}_{nP}(r_2|nr_1)\Big) \quad)\ \ ]$

$=\ \mathrm{E}\ [\ (\frac{1}{m} (\ \sum_{i=1}^{m_n} \mathrm{E}(Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}0) + \sum_{i=m_n+1}^{m} \mathrm{E}(Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}1)) \text{-}\ \mu_{\mathrm{P}1}\ )\ \ (P_{nP}(r_2|nr_1) - P_{nP}(r_2|nr_1)\ )$

$]{=}0$

$\mathcal{B}' = cov\Big(\mathrm{E}\big(\hat{P}_P(r_1) | I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\big), \mathrm{E}\big(\hat{P}_{nT}(r_2|nr_1) | I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\big)\Big)$

$= E\Big[\Big(\mathrm{E}\left(\hat{P}_P(r_1) \middle| I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\right)$

$\qquad\qquad - \mathrm{E}\big(\hat{P}_P(r_1)\big)\Big), \Big(E\left(\hat{P}_{nT}(r_2|nr_1) \middle| I(\mathrm{Y}_{1\mathrm{i}}{=}1), \mathrm{i} = 1, \dots, \mathrm{m}\right) - \mathrm{E}\Big(\hat{P}_{nT}(r_2|nr_1)\Big)\Big) \Big]$

$= \mathrm{E}\ [\ (\frac{1}{m} (\sum_{i=1}^{m_n} Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}0 + \sum_{i=m_n+1}^{m} Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}1) \text{-}\ P_P(r_1)\ )\ \ (\frac{1}{\xi m_n} \sum_{i=1}^{\xi m_n} \mathrm{E}(\mathrm{Y}_{2\mathrm{i}}^{\mathrm{nT}} |\ \mathrm{Y}_{1\mathrm{i}}{=}0) -$

$\mathrm{E}\Big(\hat{P}_{nT}(r_2|nr_1)\Big) \quad)\ \ ]$

$=\ \mathrm{E}\ [\ (\frac{1}{m} (\ \sum_{i=1}^{m_n} \mathrm{E}(Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}0) + \sum_{i=m_n+1}^{m} \mathrm{E}(Y_{1i} |\ \mathrm{Y}_{1\mathrm{i}}{=}1)) \text{-}\ \mu_{\mathrm{P}1}\ )\ \ (P_{nT}(r_2|nr_1) - P_{nT}(r_2|nr_1)\ )\ \ ]$

$= 0$

Thus $cov\big(\hat{\delta}_1, \hat{\delta}_2\big) = \mathcal{A} + \mathcal{B} - (\mathcal{A}' + \mathcal{B}') = \mathcal{A} - \mathcal{A}'{=}0$ for Binomial-Binomial scenario.

Chapter 8

Misunderstanding of a New Approach to Drug-Placebo Difference Calculation in Short Term Antidepressant-Drug Trials

(Liu, Y. (2015) Misunderstanding of a New Approach to Drug-Placebo Difference Calculation in Short Term Antidepressant-Drug Trials. *Open Journal of Statistics*, **5**, 113-119.
doi: 10.4236/ojs.2015.52014.)

Abstract: In clinical trials, drug effect is measured by a difference between subjects who are treated by experimental drug against placebo-treated subjects. In case of binary data, with observing YES/NO on each subject in certain period of time, it is the proportion of subjects who respond in treatment group minus the proportion of responders in placebo group (for example, 50% vs. 30%). However, a greater difference was proposed by Rihmer et al. (2011) with their supporting arguments, in that antidepressant response and placebo response had different mechanisms and there were equal chances for antidepressant responder to be responding to placebo and not responding to placebo at all. Therefore, the authors proposed 50% - 30% * 50% when the response rate in the treatment group and the placebo group are 50% and 30% respectively, resulting in higher drug-placebo difference than traditional understanding of 50% - 30%. In this article, we tried to explain why the authors misunderstood the drug-placebo concept for evaluating drug superiority, their misunderstanding of assumptions of traditional calculation, as well as their wrong reasoning on their proposed approach. All in all, we conclude the traditional approach of 50% - 30% is the right way of evaluating drug-placebo difference and the possible methods to control impact of placebo effect are briefly discussed at the end of this article.

Keywords: Antidepressant; Placebo Effect; Short-Term Antidepressant Effect; Unipolar Major Depression.

Section 8.1 Introduction

In clinical trials, patients are not only taking a testing drug on rigorous schedules, but also under a specific healthcare environment. Routine checks, clinical visiting and lots of psychological interviews might create a misconception to patients and clinicians and result in placebo effect. Placebo effect blunts the ability to detect drug-placebo difference in a well-controlled trail, resulting in trial failures, longer time and more resource in developing promising drugs for unmet medical needs. To deduce this trial background effects, randomization is normally applied. Subjects are randomized into either placebo or treatment group with equal probability and baseline characteristics got balanced out. With the help of randomization, only post-randomization factors and drug-placebo difference can contribute to different effects between drug and placebo groups. However, if investigators and patients have known what is given and what is taken in the trial, psychological effects will impact clinical rating scales, self-evaluation scores, compliance and patient's willingness of coordination with trial personnel. Hence blinding is essential to get rid of above impacts on evaluating drug-placebo effect. Double- blinding is a way to exclude some of those post-randomization factors. Use of placebo is to evaluate the background effect of trial procedure on patients. Placebo is sometimes better than not treated, which is seen in most psychiatry trials depending on different disease characteristics. Placebo effect is well-known in antidepressant trials. How placebo works, how placebo effect is different from drug effect, whether there are interactions between them or not, and how these issues get accounted in statistical comparison all become interesting to the academic community. And the newly proposed method on how to calculate drug-placebo difference was one particular effort to answer one aspect of these questions. What makes anti-depressant special is that general antidepressant clinical trials, especially in short-term trials, have relatively larger placebo effect than those of other drug-testing clinical trials. Section 8.2 describes complexity of placebo and antidepressant mechanisms in depressive patients. Section 8.3 evaluates drug-placebo difference under various interaction types between placebo and antidepressant responses. Section 8.4 explains all the misunderstanding of drug-placebo difference and logic errors in Rihmer et al. 2011, similar errors were also made in other two articles (Rihmer, 2007; Rihmer and Gonda 2008). Section 8.5 discusses operational management and novel designs to cope with placebo effect in antidepressant clinical trials.

Section 8.2: Mechanism of Placebo and Antidepressant Effects

Most widely used antidepressants include two classes: SSRI (selective serotonin reuptake inhibitors) and serotonin norepinephrine reuptake inhibitors. Namely, these two classes work mostly on central serotonin and norepinephrine systems (Johnson et al. 1993); Carpenter et al.,

2003) respectively. That is: the AD (antidepressant) response relies on specific underlying biological pathway in relation to biological state/illness characteristics. Moreover, due to biochemical heterogeneity, depression symptomatic improvement only occurs in certain subpopulation of individuals affected by depression. Interestingly, PL (placebo) response behaves very differently, especially from perspective of its biomarker profile. When the biomarker of change in brain glucose metabolism, a measure of positron emission tomography was monitored, PL response was shown to be associated with regional metabolic increases in the prefrontal and anterior cingulate cortices, while fluoxetine (one kind of antidepressant) response was associated with additional changes in additional changes in brainstem, striatum, and hippocampal activity (Mayberg et al., 2002). At subject level, PL (placebo) responders showed a significance increase in prefrontal cortex activity, whereas no such increase occurred in none of the rest of the population consisting of PL non-responders, AD (i.e., fluoxetine or venlafaxine) responders, and AD non-res- ponders (Leuchter et al. 2002). Moreover, most recent studies showed endogenous opiod and dopaminergic neurotransmission mediated nocebo effects, while central opioid and dopaminergic activation mediated on PL response (Enck et al. 2008); Scott et al. 2008). Then next question is how the central opoid and dopaminergic activation differs from endogenous opiod and dopaminergic neurotransmission; recent research argued that the former could mediate optimistic personality features (Sharot et al., 2007). Now the connection appears explainable, as placebo response, not with specific drug molecule, shows general response to the overall environment. For instance, some reward expectations on clinical improvement in both patients and clinicians after placebo administration, subsequently result in change in systems that mediate optimistic personality feature. So far, we can summarize that AD response and PL response work differently and could overlap in certain ways. Not everyone responds to placebo, neither does to antidepressants. From each subject, as Rihmer et al. (2011) noted patients could be divided into four different categories: (P1) AD responder and PL responder (++); (P2) AD responder and PL non-responder (+−); (P3) AD non-responder and PL responder (−+); and (P4) AD non- responder and PL non-responder (−−). All types of P1 - P4 exist in real trial results.

Section 8.3: Drug-Placebo Difference Evaluation

In this section, we would like to explore the appropriate statistical evaluation for drug-placebo difference under the circumstance of placebo response in antidepressant trials. To be more complete, let's put aside all founding in Section 8.2 first and explore all the scenarios, because some of these scenarios trigger Rihmer and co-authors (Rihmer et al. 2011) to pick up the new method over the traditional one. Therefore, it is necessary to explore all of them in detail first.

Put AD and PL response in 2X2 contingency table, then the difference between drug and placebo can be viewed marginally and jointly. Marginally means whenever we consider AD response rate, we only concentrate on AD response (response = YES and response = NO corresponding to AD = 1 and AD = 0 respectively) without considering PL mechanism. Similarly, whenever looking at PL response rate, we ignore how AD works. From Figure 8.1(a), we can clearly see that rate of response in AD group minus rate of response in PL group is first column of down diagonal minus first row of up diagonal, that is Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3. However, if we would like to look the rates jointly in terms of both AD and PL responding, then it is low left corner of down diagonal minus upper right corner of up diagonal, that is Pr(AD = 1 and PL = 0) − Pr(AD = 0 and PL = 1). Comparing to subtraction of marginal in method one in Figure 8.1(a), future specifications are needed to obtain these two joint probabilities of Pr(AD = 1 and PL = 0) − Pr (AD = 0 and PL = 1). Comparing method 1 of subtraction of marginal probabilities with subtraction of joint probabilities, we can find that they coincide with each other, since the only part in common, probability of being AD responder and PL responder, is eliminated from because residing both before the minus sign and after the minus sign. That is: Pr(AD = 1) − Pr(PL = 1) = [Pr(AD = 1 and PL = 0) + Pr(AD = 1 and PL = 1)] − [Pr(AD = 0 and PL = 1) + Pr(AD = 1 and PL = 1)] = Pr(AD = 1 and PL = 0) − Pr(AD = 0 and PL = 1). Note that, in Figure 8.2, we graphically denote divided probabilistic distribution of this joint AD and PL variables.

Assuming two difference systems mediate PL response and AD response separately, then these two systems could: (D) totally dependent; (IND) totally independent; and (Other) some dependence in between. For totally dependence, we can further divide them into 4 subcategories (Figure 8.3): (D1) all placebo responders are AD responders; (D2) all placebo responders are AD non-responders; (D3) all AD responders are placebo responders; (D4) all AD responders are placebo non-responders.

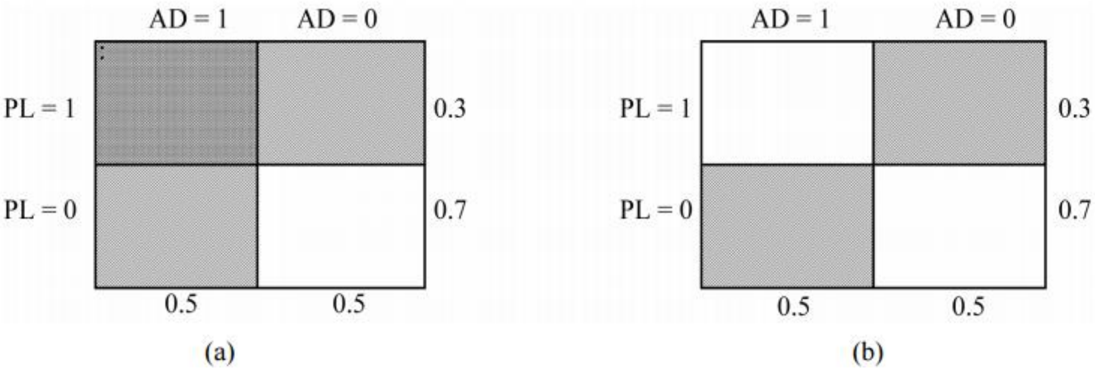


Figure 20(Fig. 8.1): Drug-placebo difference graphic representation

Figure 8.1: Drug-placebo difference graphic representation. (a) Looking at it marginally, drug-placebo difference is shaded lower diagonal minus shaded upper diagonal. (b) Looking at it jointly, drug-placebo difference is still shaded lower diagonal minus shaded upper diagonal with trellised cell deleted as compared to (a).

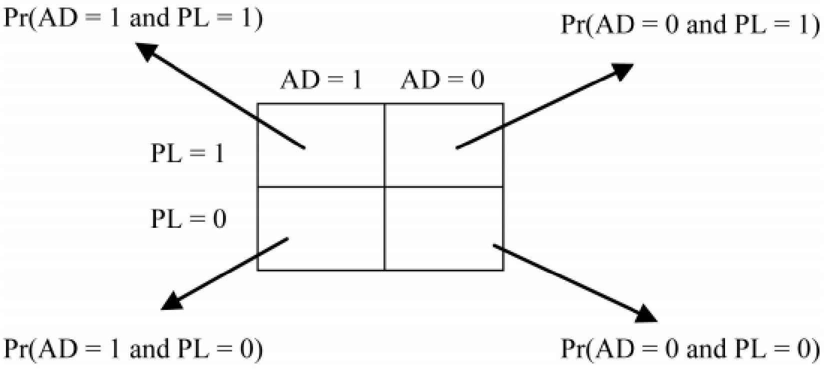


Figure 21(Fig. 8.2): Probabilistic distribution of AD/PL responses

Figure 8.2: Probabilistic distribution of AD/PL responses.

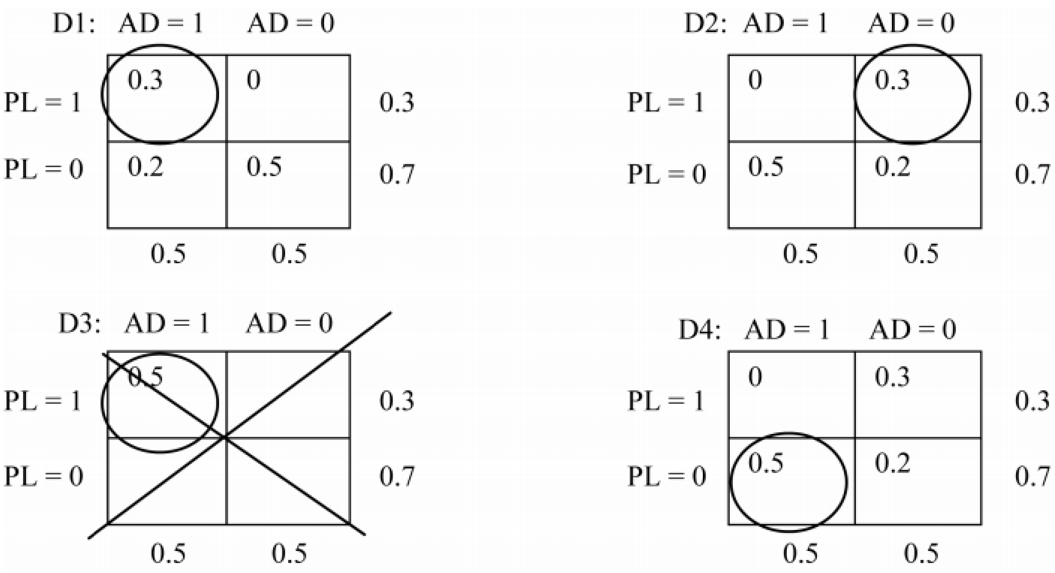


Figure 22(Fig. 8.3): Drug-placebo difference under four mutually exclusive and exhaustive scenarios

Figure 8.3: Drug-placebo difference under four mutually exclusive and exhaustive scenarios. D1: All PL responders are AD responders; D2: All PL responders are AD non-responders; D3: All AD responders are PL responders; D4: All AD responders are PL non-responders.

Section 8.3.1: Various Dependent Structures

(D1): Dependence scenario 1. Since all PL responders are AD responders, Pr(AD = 1|PL = 1) = 1. Circled cell Pr(AD = 1 and PL = 1) = Pr(AD = 1|PL = 1) * Pr(PL = 1) = 1 * 0.3 = 0.3; and then drug-placebo difference = Pr(AD = 1 and PL = 0) − Pr(AD = 0 and PL = 1) = 0.2 − 0 = 0.2 = Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3.

(D2): Dependence scenario 2. Since all PL responders are AD non-responders, Pr(AD = 0|PL = 1) = 1. Circled cell Pr(AD = 0 and PL = 1) = Pr(AD = 0|PL = 1) * Pr(PL = 1) = 1 * 0.3 = 0.3 and drug-placebo difference = Pr(AD = 1 and PL = 0) − Pr(AD = 0 and PL = 1) = 0.5 − 0.3 = Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3.

(D3): Dependence scenario 3. Intuitively, this can't exist because: if all AD responders are PL responders, PL responder rate will be greater or equal to AD responder rate, which contradicts our assumption of probability of AD equal to 1 being 0.5 and PL equal to 1 being 0.3 respectively. Had we have PL responder rate exceeded AD responder rate; this would be a wrong target drug to develop since its effect is numerically inferior to placebo. Mathematically, if we have all AD responders are PL responders, conditionally probability of Pr(PL = 1|AD = 1) = 1. Therefore, Pr(AD = 1 and PL = 1) = Pr(PL = 1|AD = 1) * Pr(AD = 1) = 1 * 0.5 > Pr(PL = 1) = 0.3. This violates probability axiom, as Pr(PL = 1) = Pr(AD = 1 and PL = 1) + Pr(AD = 0 and PL = 1) and should not be less than Pr(AD = 1 and PL = 1) alone. This calculation proves our intuitive interpretation: under the condition of all AD responders are PL responders, existing of AD non-responders being PL responders will lead to greater PL response rate than AD response rate, in which is against the goal of drug development.

(D4): Dependence scenario 4. Since all AD responders are PL non-responders, Pr(PL = 0|PL = 1) = 1. Circled cell Pr(PL = 0 and AD = 1) = Pr(PL = 0|AD = 1) * Pr(AD = 1) = 1 * 0.5 = 0.5 and drug-placebo difference = Pr(AD = 1 and PL = 0) − Pr(AD = 0 and PL = 0) = 0.5 − 0.3 = 0.2 = Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3. Graphically, dependence scenario 2 equals dependence scenario 4. Let's try to prove it mathematically.

Claim: D2 dependence structure is the same as D4 dependence structure.

Proof: D2 = >D4

Pr(AD = 1 and PL = 1) + Pr(AD = 1 and PL = 0) + Pr(AD = 0 and PL = 1) + Pr(AD = 0 and PL = 0) = 1

$\Rightarrow$ Pr(AD = 1 and PL = 1) + Pr(PL = 0|AD = 1) * Pr(AD = 1) + Pr(AD = 0|PL = 1) * Pr(PL = 1) + Pr(AD = 0 and PL = 0) = 1

Because Pr(AD = 0|PL = 1) = 1, then Pr(PL = 0|AD = 1) * Pr(AD = 1) = 1 − 1 * Pr(PL = 1) − Pr(AD = 1 and PL = 1) − Pr(AD = 0 and PL = 0) = Pr(PL = 0) − Pr(AD = 1 and PL = 1) − Pr(AD = 0 and PL = 0)

= Pr(AD = 1 and PL = 0) − Pr(AD = 1 and PL = 1)

= Pr(AD = 1) * Pr(PL = 0|AD = 1) − Pr(AD = 1) * Pr(PL = 1|AD = 1)

After Canceling Pr(AD = 1) from both sides, we have Pr(PL = 0|AD = 1) = Pr(PL = 0|AD = 1) − Pr(PL = 1|AD = 1)

$\Rightarrow$ Pr(PL = 1|AD = 1) = 0

$$\Rightarrow \frac{\Pr(\mathrm{PL} = 1 \text{ and } \mathrm{AD} = 1)}{\Pr(\mathrm{AD} = 1)} = 0$$

$\Rightarrow$ Pr(PL = 1 and AD = 1) = 0

Together with Pr(PL = 0 and AD = 1) + Pr(PL = 1 and AD = 1) = Pr(AD = 1)

$\Rightarrow$ Pr(PL = 0 and AD = 1) = Pr(AD = 1)

$\Rightarrow$ Pr(PL = 0|AD = 1) * Pr(AD = 1) = Pr(AD = 1)

$\Rightarrow$ Pr(PL = 0|AD = 1) = 1, because Pr(AD = 1) is a positive number.

Pr(PL = 0|AD = 1) = 1 is for D4 structure. All AD responders are PL non-responders. □

Similarly, we can show D4 => D2.

In summary, under all reasonable dependence scenarios (i.e., D1 - D4 excluding D3), 4 cell probabilities are fixed and drug-placebo difference using joint probabilities is available. However, as discussed in Section 8.2, this drug-placebo difference is always 0.5 − 0.3, the same as that of being obtained by marginal probabilities. The other reason to have detailed discussion about above mutually exclusive and exhaustive scenarios is for later discussion about the method proposed by Rihmer et al. (2011).

Section 8.3.2: Independent Structure

If the mechanism of placebo response is independent of that of antidepressant response, placebo responders can randomly either to be AD responder or to be AD non-responder. Similarly, AD responders have an equal chance to either be PL responder or be PL non-responder. Being a placebo responder is independent of being an AD responder. Then, under this scenario, what about drug-placebo difference? In Figure 8.4, we see that since Pr(AD = 1|PL = 1) = 0.5, we have Pr(AD = 1 and PL = 1) = Pr(AD = 1|PL = 1) * Pr(PL = 1) = 0.5 * 0.3 = 0.15. Then drug-placebo difference using joint probability is 0.35 − 0.15 = 0.2, numerically exactly the same as Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3 = 0.2 using marginal probabilities.

Section 8.3.3: Structures between Totally Dependent and Totally Independent

If neither definite dependence nor independence presents, some other structures in between play a role for mechanisms of placebo and AD responding. As in the 2X2 contingency table (Figure 8.2), once one cell probability is fixed, all other cells are known as well. For instance, probability of both AD and PL (i.e., Pr(AD = 1 and PL = 1)) responding is known. In example 1, with Pr(AD = 1 and PL = 1) = 0.25 known (bigger than the probability under independence in Figure 8.4), drug-placebo difference can be calculated as Pr(AD = 1 and PL = 0) = Pr(AD = 0 and PL = 1) = 0.25 − 0.05 = 0.2, the same as Pr(AD = 1) − Pr(PL = 1) = 0.5 − 0.3 = 0.2. In example 2, with Pr(AD = 1 and PL = 1) = 0.1 known (smaller than its probability under independence scenario), drug-placebo difference can be calculated as Pr(AD = 1 and PL = 0) = Pr(AD = 0 and PL = 1) = 0.4 − 0.2 = 0.2. As shown in Figure 8.5, Pr(AD = 1 and PL = 1) can be either greater than that of independence scenario in example 1, or less than that of example 2. No matter it is higher or lower than that of independence structure, once joint probabilities are known, drug-placebo difference can easily derived, which again is the same as the marginal probability difference. The advantage of using marginal probability is that joint probabilities are normally unknown due to unobservable property and can't be used to derived drug-placebo difference. On the contrary, marginal probabilities are always observable and hence can easily be used for evaluating drug superiority.

In clinical trials, we measure response on each subject, and group them into treatment versus placebo to find a measure so that superiority of drug vs. placebo can be evaluated and tested. Each joint probability is actually unobservable in the trial except under wholly independence or dependence structures. It may be possible to use another trial to test independence assumption, but normally we can just reject or fail to reject independence hypothesis. Still, we can't prove it is indeed independent. For dependence structure, even with an external trial specifically for evaluating dependence structure, it is really hard to prove which dependence structure it is. Also, from Section 8.2, the presence of AD non-responder and PL responders excludes the possibility of having dependence scenario 1, which is all PL responders are AD responders; similarly, the presence of AD responders and PL responders excludes dependence scenarios 2 and 4, which are all PL responders are AD non-responders and all AD responders are PL non-responders respectively.

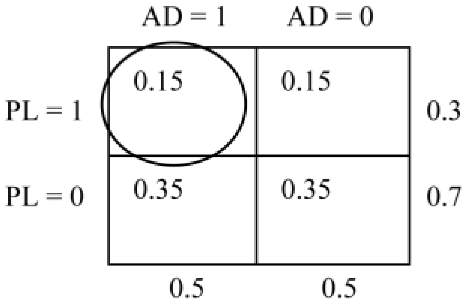
From general discussion in Section 8.2 and each specific example in Section 8.3, we all show that drug-placebo difference can be evaluated by marginal probability difference.

Section 8.4: Discussion of Misunderstanding Leading to a Wrong New Approach

After stating and proving the right way of evaluating drug-placebo difference, we now have to discuss why the proposed method by Rihmer et al. (2011) is wrong and where the logic flaws resided in their article. There are several steps for Rihmer et al. (2011) to propose $0.5 - 0.3 *$

50% and reason against the traditional method of 0.5 − 0.3. First of all, they thought that old method of 0.5 − 0.3 depends on the assumption of all PL responders being AD responders (i.e., Pr(AD = 1|PL = 1) = 1), which corresponds to dependence structure 1 in Figure 8.3. This is indeed wrong. Under dependence structure 1, Then the authors had a wrong perspective that drug-placebo difference is Pr(AD = 1 and PL = 0) = Pr(AD = 1) − Pr(AD = 1 and PL = 1) = 0.5 − 0.3 using joint probabilities in Figure 8.3 Dependence 1 table. This is actually using a wrong rational but to end up with a correct number of 0.2. Later they thought that more consideration should be put into Pr(AD = 1 and PL = 1) to account for the fact that not all PL responders can be AD responders. Under independence structure, there is equal probability for a PL responder to be an AD responder or not to be an AD responder. Hence they went to independence structure in Figure 8.4. As joint probabilities in Figure 8.4 show, Pr(AD = 1 and PL = 0) = Pr(AD = 1) − Pr(AD = 1 and PL = 1) = 0.5 − 0.15 = 0.35. We think that Rihmer and co-authors [1] started with wrong assumptions for drug-placebo difference; used wrong measure for it; had a wrong interpretation for this measure; and subsequently proposed a wrong approach. Now, let explain further about why probability of being an AD responder but not a PL responder (i.e., Pr(AD = 1 and PL = 0)) is not a right measure of drug-placebo difference. This measure is measuring the chance for each individual to be AD responder and PL non-responder simultaneously; or is measuring relative frequency of subjects who are AD responder but not PL responder in the whole population. Either interpretation has nothing to do with the drug-placebo difference, which is the relative frequency of AD responders over PL responders in antidepressant patient population. And this joint probability is normally unobservable in the clinical trials, where patients are randomly assigned to PL or AD to obtain efficacy measure to assess AD relative superiority. On the contrary, each patient is a unit to be treated by either placebo or AD;

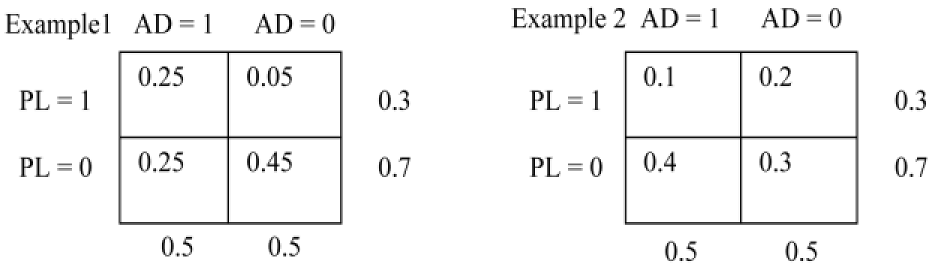
responder rate in AD-treated group minus the responder rate in the PL-treated group provide an objective measure for drug-placebo difference after all baseline factors being balanced out by randomization and the only factor contributing to drug-placebo difference is what they have received in the trial. This, as shown in Section 8.3, is irrespective of what kind of joint mechanism between drug and placebo responses. Besides, calculation from marginal rate difference is the same as calculating difference from joint probabilities, whereas the latter is normally unobservable and can't be obtained from this randomized clinical trial.



	AD = 1	AD = 0	
PL = 1	0.15	0.15	0.3
PL = 0	0.35	0.35	0.7
	0.5	0.5	

Figure 23(Fig. 8.4): Drug-placebo difference under independent structure

Figure 8.4: Drug-placebo difference under independent structure



Example1	AD = 1	AD = 0	
PL = 1	0.25	0.05	0.3
PL = 0	0.25	0.45	0.7
	0.5	0.5	

Example 2	AD = 1	AD = 0	
PL = 1	0.1	0.2	0.3
PL = 0	0.4	0.3	0.7
	0.5	0.5	

Figure 24(Fig. 8.5): Two examples of drug-placebo difference under structures between totally dependent and independent

Figure 8.5: Two examples of drug-placebo difference under structures between totally dependent and independent. Example 1: probability of being AD and PL responders is greater than that of independence structure; Example 2: probability of being AD and PL responders is lower than that of independence structure.

Section 8.5: Discussion of Operational Management and Novel Designs to Cope with Placebo Effect in Antidepressant Clinical Trials

After the discussion of the right way of understanding and evaluating drug-placebo difference and pointing out all the flaws in Rihmer and co-authors’ wrong proposal, it seems that we are going back to the original place to favor traditional method of $\Pr(AD = 1) - \Pr(PL = 1)$. Then what should we do to avoid jeopardizing a trial because of placebo effect? And should we just let it go unchecked? Of course, the answer is no. This is actually a very interesting but complicated area and not intended to be covered in this article. Here, we can briefly point out some related perspectives. To avoid failure trial due to placebo effect, we can put more efforts on innovated design and manage it more appropriate in operation. The main challenge is to lower the optimistic expectation from both patient and clinician. Since higher placebo response was found in mild-moderate depression, excluding these patients in the trial should be considered. And more scientific scoring system, more self-scoring scale, help from biomarker markers, and/or central rating could be combined to narrow the possibility of overstated expectation. Mathematically, novel designs as sequential parallel designs are also available in the literature.

References

Rihmer, Z., Gonda, X., Döme1, P., Erdős, P., Ormos, M. and Pani, L. (2011) Novel Approaches to Drug-Placebo Difference Calculation: Evidence from Short-Term Antidepressant Drug-Trials. *Human Psychopharmacology: Clinical and Experimental*, 26, 307-312.

Rihmer, Z. (2007) Drug-Placebo Difference: In Antidepressant Drug Trials Could Be 50% Greater Than Previously Believed. *Neuropsychopharmacologia Hungarica*, 9, 35-37.

Rihmer, Z. and Gonda, X. (2008) Is Drug-Placebo Difference in Short-Term Antidepressant Drug Trials on Unipolar Major Depression Much Greater Than Previously Believed? *Journal of Affective Disorders*, 108, 195-198.http://dx.doi.org/10.1016/j.jad.2008.01.020

Johnson, M.R., Lydiard, R.B., Morton, W.A., Laird, L.K., Steele, T.E., Kellner, C.H., et al. (1993) Effect of Fluvoxamine, Imipramine and Placebo on Catecholamine Function in Depressed Outpatients. *Journal of Psychiatric Research*, 27, 161-172.http://dx.doi.org/10.1016/0022-3956(93)90004-L

Carpenter, L.L., Anderson, G.M., Siniscalchi, J.M., Chappell, P.B. and Price, L.H. (2003) Acute Changes in Cerebrospinal Fluid 5-HIAA Following Oral Paroxetine Challenge in Healthy Humans. *Neuropsychopharmacology*, 28, 339- 347.http://dx.doi.org/10.1038/sj.npp.1300025

Mayberg, H.S., Silva, J.A., Brannan, S.K., Tekell, J.T., Mahurin, R.K., McGinnis, S., et al. (2002) The Functional Neuroanatomy of the Placebo Effect. *The American Journal of Psychiatry*, 159, 728-737. http://dx.doi.org/10.1176/appi.ajp.159.5.728

Leuchter, A.F., Cook, I.A., Witte, E.A., Morgan, M. and Abrams, M. (2002) Changes in Brain Function of Depressed Subjects during Treatment with Placebo. *The American Journal of Psychiatry*, 159, 122-129. http://dx.doi.org/10.1176/appi.ajp.159.1.122

Enck, P., Benedetti, F. and Schedlowski, M. (2008) New Insights into the Placebo and Nocebo Responses. *Neuron*, 59, 195-206.http://dx.doi.org/10.1016/j.neuron.2008.06.030

Scott, D.J., Stohler, C.S., Egnatuk, C.M., Wang, H., Koeppe, R.A. and Zubieta, J.K. (2008) Placebo and Nocebo Effects Are Defined by Opposite Opioid and Dopaminergic Responses. *Archives of General Psychiatry*, 65, 220-231.http://dx.doi.org/10.1001/archgenpsychiatry.2007.34

Sharot, T., Riccardi, A.M., Raio, C.M. and Phelps, E.A. (2007) Neural Mechanisms Mediating Optimism Bias. *Nature*, 450, 102-105. http://dx.doi.org/10.1038/nature06280

Chapter 9

Optimal Group Sequential Designs Constrained on both Overall and Stage One Error Rates

(to be submitted)

Abstract: Optimized group sequential designs proposed in the literature have designs minimizing average sample size (ASN) with respect to a prior distribution of treatment effect with overall type I and type II error rates well-controlled. The optimized asymmetric group sequential designs that we present here additionally consider constrains on stopping probabilities at stage one: probability of stopping for futility at stage one when no drug effect exists as well as the probability of rejection when the maximum effect size is true at stage one so that accountability of group sequential design is ensured from the very first stage throughout. Besides, non-binding efficacy bounds are used to account for often-occurred overrunning in real trials, and the shape parameters for Wang-Tsiatis upper bounds and Kim-DeMets lower bounds are utilized to find optimized group sequential designs minimizing ASN while maintaining error and power requirements overall and at stage one. From examples illustrated, the maximum sample size determined through optimization turns out to be smaller than prior optimized designs using other ways of optimization.

Keywords: Group sequential design; Optimization; Asymmetric; Non-binding; Overrunning.

Section 9.1: Introduction

After publication of computational work by Armitage, McPherson and Rowe (1969), research on group sequential tests have been proposed including those of Haybittle (1971), Peto et al., (1976), Pocock (1977), O’Brien and Fleming (1979), Harrington and O’Brien (1984) and Wang and Tsiatis (1987). “Error spending function” introduced by Lan and DeMets (1983) allows more flexibility in group sequential designs when the number of stages is unpredictable at trial start or interim analysis is delayed past the planned timing of analyses as a trial proceeds so that boundaries need to be adjusted during the course of a trial. Jennison (1987) derived optimal one-sided group sequential tests concerning the mean of a normal distribution with known variance, which are optimal in that expected sample size is minimized under given values of the mean or averaged over several values of the mean subject to constraints on the overall type I and type II error probabilities. Using backward algorithm, Eales and Jennison (1992) and Eales (1995) derived optimal group sequential tests for one-sided and two-sided scenarios, respectively.

Backward algorithm, though being one-dimensional, is quite complicated to implement. Anderson (2007) made use of shape parameters of overall type I and II error spending functions to derive optimized group sequential tests that minimize expected sample size or expected squared sample size , which lessens computational load compared with the backward algorithm proposed by Eales and Jennison (1992) and Eales (1995). Previous optimized group sequential tests are all subject to constraints on overall type I and type II error probabilities only. The method we present in this article additionally considers stopping probabilities at the first interim analysis when the maximum effect size is true or to stop for futility at stage one when the null hypothesis is true. Controlling probability at stage one is essential when the rejection/acceptance conclusion can be drawn at stage one, which is unfortunately ignored in many published optimal group sequential procedures. Section 9.2 builds up the basics (i.e., notation and other preliminaries). Section 9.3 illustrates how optimization is done. Section 9.4 shows the results for optimized asymmetric group sequential tests with respect to desired prior distribution of the parameter of interest. Section 9.5 discusses features of proposed optimized designs compared with prior optimized designs. For example, the one proposed by Anderson (2007).

Section 9.2: Notations

Section 9.2.1: A Motivating Example

For a trial with survival end point of time to relapse/death/failure, an event such as a relapse/death/failure in the randomization phase is defined as meeting one of the criteria for the first time after randomization. The objective of the trial is to test the superiority of drug against placebo in delaying time to relapse as an example from now on in the randomization phase after randomization with efficacy summarized by effect size δ, log hazard ratio divided by its variance. Detailed information is summarized in Table 9.1. There is a 50% of chance to have an effect size (standardized log hazard ratio of placebo relative to drug) equal to zero (i.e., under null hypothesis). There is a 50% of chance to have an effective drug (i.e., under alternative hypothesis); the conditional probability of having the optimal effect size is 20% (standardized log hazard ratio equal to 0.755 and relapse rate being 35% and 60% for drug and placebo, respectively); the conditional probability of having the expected effect size is 20% (standardized log hazard ratio (placebo vs. drug) = 0.617 and relapse rate being 35% and 55% for drug and placebo, respectively); and the conditional probability of having minimal effect size of interest is 50% (standardized log hazard ratio being 0.476 and relapse rate being 35% and 50% for drug and placebo, respectively). A design is preferred to incorporate all information regarding the prior information on effect δ. Hence an optimized group sequential design for minimizing average sample number while subject to a set of constrains needs to be developed. In order to ensure power and that the false positive rate to be well-controlled not only in the overall sense but also for every single interim analysis, we have to control error probabilities at stage one. Since tests in group sequential designs use cumulative data up to the testing stage, controlling error probabilities at stage one can guarantee validity of tests at subsequent stages. Inspired by design specifications i) and ii) on Page 141 of Liu and Chi (2001) and controlling of probability of continuing to later stages when the null hypothesis is true at stage one in Liu et al (2012), our optimized group sequential designs are constructed to ensure sufficient power to reject the null under $\delta_{\max}$ (the maximum effect size) even at stage one and a proper probability for stopping for futility at Stage One if null is true; and the overall power is calculated under the minimal effect size $\delta_{\min}$ instead of expected effect size to avoid resulting in an underpowered study when true effect size is in between $\delta_{\min}$ and the expected effect size and the whole trial was prospectively powered under the expected effect size. More specifically, the optimized design has the following operational properties:

1) the power of rejecting the null hypothesis H_0 based on data from stage one is at least 1- β, say 0.8 or 0.9, if the true effect size is $\delta_{\max}$ (i.e., 5.15 in our example);

2) the overall power to reject the null hypothesis H_0 is at least 1- β, if the true effect size is $\delta_{\min}$ (i.e., 3.24 in our example);

3) the overall type I error rate (one-sided) to reject null H_0 is α, say 0.025;

4) if H_0 is true, the probability of continuing to stage two while not stopping for futility at stage one is at most α_F, say 0.3 or 0.2; and

5) non-binding upper efficacy boundaries are employed to account for overrunning data.

Table 34(Tab. 9.1): Knowledge of relative effectiveness of drug and placebo prior to trial start

Table 9.1: Knowledge of relative effectiveness of drug and placebo prior to trial start, with 'logHR' means log of hazard ratio.

Hypothesis (Probability)	Conditional probability	Difference in relapse rates (Placebo-drug)	Relapse rate		# of events needed (fixed sample design)	Effect size= Log(Hazard ratio(Placebo/drug))/ $\sqrt{(4/\# \ of \ events)}$
			drug	placebo		
H_A (50%)	20%	25% (optimal or maximum)	35%	60%	74	logHR=0.755, δ_{opt}=δ_{max}=5.15
	30%	20% (expected)	35%	55%	111	logHR=0.617, δ_{exp}=4.21
	50%	15% (minimal)	35%	50%	186	logHR=0.476, δ_{min}=3.24
H_0 (50%)	100%	0%	NA	NA	NA	logHR=0, Effect size=0

Section 9.2.2 Group Sequential Setting

Considering a group sequential trial with K planned analyses, let δ be the parameter of interest, a measure of placebo-drug difference and assume it can be estimated from trial data. The distribution of statistics Z_1 , Z_2 , …, Z_K are derived from cumulative data up to stages from 1, 2 …, K, and follows a canonical joint form (Chapter 3, Jennison and Turnbull (2000)) of multivariate normal distribution with $E(Z_i) = \delta\sqrt{I_i}$ and $Cov(Z_i, Z_j) = \sqrt{I_i/I_j}$, $1 \leq i \leq j \leq K$ and $\{I_1, \dots, I_K\}$ are information levels for parameter δ. For the motivating example described above, the standardized log-rank statistic (Chapter 13.2, Jennison and Turnbull) approximately has the canonical joint distribution, given information level I_i proportional to the number of events at the ith interim analysis.

Section 9.2.3 Non-binding Efficacy Upper Boundaries

When a group sequential test is proposed t to test the null hypothesis $H_0 = 0$ against $H_A = \delta$ for fixed $\delta > 0$ with overall probability of rejecting null at most α, say 0.025 for one-sided test when null hypothesis is true, and overall probability of rejecting null with power of 1- β when the alternative hypothesis is true and the drug is effective, the null hypothesis will be rejected at stage i when the observed statistic $Z_i \geq u_i$ or trial is stopped early for futility if $Z_i \leq l_i$, where l_i and u_i are, respectively, the stage i lower futility and upper efficacy boundaries. During the trial, it takes time to close a site and then re-open it, or initiate new sites. At the time of interim analysis, without knowing the trial results and not knowing if the trial should be stopped or not, sites normally continue recruiting new subjects or subjects remained event-free are kept being treated during the period of conducting interim analysis. If the stopping for trial for efficacy or for futility can be claimed by interim data, overrunning data occurred succeeding interim cutoff date is inevitably accumulated. Based on the intent-to-treat principle, all randomized subjects should be included in the analysis because randomization is supposed to balance out impact of baseline characteristics on treatment effect and the final analysis including complete data should be conducted and included in the submission document per regulatory requirement. This practical issue poses some requirements on choosing a proper group sequential design as explained below.

Binding upper efficacy bounds namely indicate that upper bounds are derived under the consideration of lower bounds while otherwise not being considered for non-binding efficacy bounds. If the interim analysis suggests stopping for efficacy at interim, conducting final analysis including overrunning data will not inflate type I error rate regardless of whether upper efficacy boundaries are binding or not binding with lower bounds because the drug has been shown to be effective at interim and one more rejection on the same hypothesis won't impact type I error rate; however, if the interim analysis shows stopping for futility, binding upper efficacy boundaries might inflate overall type I error rate because rejecting null at final analysis with futility bound crossed earlier on is not considered at all originally. In this case, non-binding efficacy boundaries can solve this dilemma, in which lower bounds are ignored when deriving upper efficacy boundaries and the null hypothesis may be rejected at final analysis, even though the trial has had futility criterion $Z_i \leq l_i$ met at interim.

Section 9.2.4 Wang-Tsiatis Family as Upper Boundaries and Kim-Demets Family as Lower Boundaries

Group sequential tests allow stopping the trial and rejecting the null hypothesis at stage i when the observed statistic $Z_i \geq u_i$ or stopping and accepting the null and stopping for futility if $Z_i \leq l_i$. Wang and Tsiatis (1987) proposed a family of boundary function of the form

$$u_i = (k/K)^{\rho-1/2}C \qquad (9.1)$$

where the shape parameter $\rho \in (-\infty, +\infty)$, k = 1, 2,…, K, and C is a constant. It is known that this family gives a Pocock boundary when $\rho = \frac{1}{2}$ and an O'Brien-Fleming boundary when $\rho = 0$. Liu and Anderson (2008a, 2008b) proposed using sequential p-value to obtain inference after group sequential test considering the totality of data; and they argued that sequential p-value with help from the Wang-Tsiatis boundary function, compared with other boundary functions, has special inferential meaning because it connects to the maximum likelihood estimate of δ, directed likelihood statistic and score statistic when ρ equals 1, ½ and 0, respectively (Section 3.1, Liu and Anderson (2008b)). The special inferential meaning carried by Wang-Tsiatis (referred to as 'WT') also made us use it as the upper boundary function to search for optimized tests, in which Wang-Tsiatis' shape parameter plays an important role in optimization.

Once upper bounaries are defined, Kim and DeMets (1987) (referred to as 'KD') $\beta -$spending function can be used to find lower boundaries which ensure a certain power to be achieved under the alternative hypothesis. For $i = 1,2 \ldots, K$, the type II error spent at stage i is denoted as

$$\beta_i(\delta_{\min}) = P_{\delta_{\min}}\left\{\{Z_i \le l_i\} \cap_{j=1}^{i-1} \{l_j \le Z_j \le u_j\}\right\} \qquad (9.2)$$

and then summing over stages, $\beta(\delta_{\min}) = \sum_{j=1}^{K} \beta_j(\delta_{\min})$ results in the overall type II errror, which is the desired probability of crossing lower boundary at any analysis when $\delta_{\min}$ is the true value for parameter of interest, δ.

We wish to set lower boundary l_i to obtain $\beta(\frac{I_i}{I_K}, \delta_{\min})= \sum_{j=1}^{i} \beta_j(\delta_{\min})$, where on the other hand accumulating type II error up to stage i $\beta(\frac{I_i}{I_K}, \delta_{\min})$ is determined by $\beta(\frac{I_i}{I_K})^{\gamma}$ using Kim-DeMets function. That is: the Kim-DeMets function of $\beta(\frac{I_i}{I_K})^{\gamma}$ determines the cumulative type II error up to Stage i , and then we use Equation 9.2 to back calculate lower bounds $\{l_1, \ldots l_K\}$ and information level vectors $\{I_1, \ldots I_K\}$ to achieve the required overall power.

Section 9.2.5: Operational Characteristics of Proposed Optimized Group Sequential Design

Shape parameters ρ and γ mentioned above in Section 9.2.4 play a very important role in finding optimized group sequential designs to accommodate Criterion 1-5 in Section 9.2.1, whereby these 5 requirements can mathematically be formulated as follows:

$$P_0\{Z_1 \geq l_1 \cup Z_2 \geq u_2 \cup, \cdots, \cup Z_K \geq u_K\} = \alpha \qquad (9.3)$$

$$P_{\delta_{max}}\{Z_1 \geq u_1\} \geq 1 - \beta \qquad (9.4)$$

$$P_{\delta_{min}}\{Z_1 \leq u_1\} + P_{\delta_{min}}\{l_1 \leq Z_1 \leq u_1, Z_2 \geq u_2\} + \cdots + P_{\delta_{min}}\{l_1 \leq Z_1 \leq u_1, \ldots, l_{K-1} \leq Z_{K-1} \leq u_{K-1,}\, Z_K \geq u_K\} = \beta \qquad (9.5)$$

$$P_0\{Z_1 \geq l_1\} = \alpha_F \qquad (9.6)$$

The requirement for overall type I error control with non-binding upper bounds is described in Equation 9.3; overall type II error (or power) requirement is depicted in Equation 9.5; first stage requirement for power to stop for efficacy when the maximum effect size is true is in Equation 9.4; and the stop for futility at stage one when there is no effect at all is clearly stated in Equation 9.6. The way how error rates in stage one are controlled is illustrated in the optimization steps below (Section 9.3.2). Appendix 9.1 shows that we can always find information time pint t_1 to ensure large enough probability of rejecting for efficacy under maximum effect size in our proposed algorithm.

On the contrary, Anderson (2007) and other publications on optimized group sequential designs only considered overall type I (Equation 9.3) and type II error rate (Equation 9.5) without considering stage one probabilities (Equations 9.4 and 9.6). Additional considerations on stage one error rates in Equations 9.4 and 9.6 further ensure proper design features starting from stage one and throughout. Furthermore, the whole trial is powered at the minimal effect size δ_{min} in our consideration to be more conservative and to avoid an underpowered study in case the true effect size is in between δ_{min} and the expected effect size while the whole trial was erroneously powered under the expected effect size.

Section 9.3: Optimization

Section 9.3.1: Objective Function for Optimization

After finding $3K$ parameters of a particular group sequential design, $\{u_1, \ldots u_K\}$, $\{l_1, \ldots l_K\}$ and $\{I_1, \ldots I_K\}$, expected sample number, denoted as $E_\delta(n)$, at a particular alternative can be computed (P237, Jennison and Turnbull (2000)). From Table 9.1, we know the prior distribution of δ is: 50% chance of being 0, 10% chance of being maximum/optimum effect size of $\delta_{max} = 5.15$, 15% chance of being at expected effect size of $\delta_{exp} = 4.21$ and 25% chance of being minimum effect size $\delta_{min} = 3.24$. Our objective function to minimize is average of $E_\delta(n)$ with respect to prior distribution of δ. That is ASN = $\sum_{\delta \in M} E_\delta(n)P(\delta)$, where M is the range of δ and we have four options for δ in our motivating example.

Section 9.3.2: Optimization Strategy And Numerical Calculation

When the shape parameter for Wang-Tiastis family function, ρ, is given, ASN increases as α_F decreases. In order to minimize ASN, null probability of failure to stop at stage one is chosen, say α_F=0.3. That is: when there is no effect for testing drug, the probability of stopping for futility at stage one is 0.7 (i.e., 1 minus 0.3). Figure 9.1 illustrates some points of the proposed optimization strategy.

Step 1: For a given standardized information vector t (with first stage information fraction t_1 together with equally spaced remaining stages), type I error α and a shape parameter ρ for Wang-Tiastis function, upper bounds $\{u_1, \ldots, u_K\}$ are then obtained.

Step 2: Given α_F , for example 0.3, together with t vector, α, β and ρ, Kim-DeMets shape parameter γ is chosen so that overall power under δ_{min} is 1- β and the the probability of continuing to stage two is 1-α_F. In this step, lower boundaries $\{l_1, \ldots\ l_K\}$ and information vector $\{I_1, \ldots, I_K\}$ are determined. Now γ is a function of α, α_F, β and ρ, denoted as $\gamma_{(t_1, \alpha,\ \alpha_F, \rho, \beta)}$.

Step 3: Check if $P_{\delta_{\max}}\{Z_1 \geq u_1\} \geq 1-\beta$ (Equation 9.4) is met. If not, increase information level spent at stage one (i.e., t_1) and then redefine t vector with new t_1 and equally spaced remaining stages, repeat Steps 1 and 2 until Equation 9.4 is met (Appendix 9.1).

Step 4: Repeat Steps 1-3 for a range of values of ρ, for example ρ_1, ρ_2, ρ_3, ..., and find the ρ^* which gives minimal value of ASN with respect to prior distribution of δ while comforming to Criteria 1-5 in Section 9.2.1.

Step 5: After finding ρ^*, pick up $\gamma_{(t_1,\alpha,\,\alpha_F,\,\rho^*,\beta)}$ which is the lower shape parameter to make the design meet Criteria 1-5 and based on ρ^*.

Step 6: For a given set of α, α_F, β, t_1 and searched pair of optimal shape parameters $(\rho^*, \gamma_{(t_1,\alpha,\,\alpha_F,\rho^*,\beta)})$, output optimized design with 3K parameters of $\{l_1, \dots\ l_K\}$, $\{u_1, \dots, u_K\}$, $\{I_1, \dots, I_K\}$ and corresponding operational characteristics based on chosen optimal shape parameters.

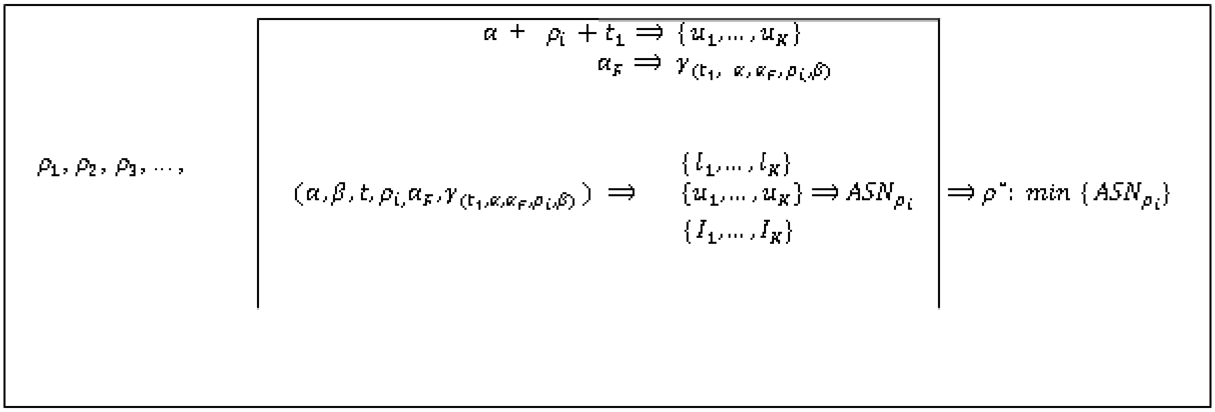

$\alpha + \rho_i + t_1 \Rightarrow \{u_1, \dots, u_K\}$
$\alpha_F \Rightarrow \gamma_{(t_1,\ \alpha,\alpha_F,\rho_i,\beta)}$
$\rho_1, \rho_2, \rho_3, \dots,$
$(\alpha, \beta, t, \rho_i, \alpha_F, \gamma_{(t_1,\alpha,\alpha_F,\rho_i,\beta)}) \Rightarrow$
$\{l_1, \dots, l_K\}$
$\{u_1, \dots, u_K\} \Rightarrow ASN_{\rho_i}$
$\{I_1, \dots, I_K\}$
$\Rightarrow \rho^*: min\ \{ASN_{\rho_i}\}$

Figure 25(Fig. 9.1): Graphic illustration

Figure 9.1: Graphic illustration of optimization using shape parameter ρ and γ.

Seen from Figure 9.1 and optimization steps, upper bounds can be determined by overall type I error, standard information vector t and a WT shape parameter ρ. Subsequently, upper bounds together with overall type II error and stage one futility error α_F to make sure probability of continuing into stage two under null being 1-α_F, lower KD shape parameter γ can be searched to fullfil given requirements. Now a specific group sequential design is defined, and probability of rejection at stage one under maximum effect size is then checked to make sure this probability is also 1- β (Equation 9.4). If not, standard information vector can be re-defined to have a larger t_1(Appendix 9.1) along with equally spaced subsequent stages and then re-do all previous steps to set corresponding lower shape parameter γ together with upper/ lower bounds. Finally, in the space of shape parameter of ρ, a spectrum of group sequential designs can be defined so that ρ^* that minimizes ASN with regards to the prior distribution of effect size can be explicitly sought out. In the end, we have optimal upper shape parameter ρ^*, corresponding $\gamma_{(t_1,\alpha,\alpha_F,\rho^*,\beta)}$ and all other operation characterisitics for this optimal design. In all our examples below, we start with $t_1 = 0.5$, which already meets the criterion of stopping for efficacy under maximum effect size with probability greater than $1 - \beta$ (Equation 9.4). Therefore, no further increase of t_1 is needed.

One question that still remains unclear is: how would we iteratively find the information vector $\{I_1, \ldots, I_K\}$ in Step 2? The trick is to set the a standardized information vector $\{t_1, \ldots, t_K\}$ with t_K=1 first (for example: K=10, we have t={0.5, 0.55,0.61,0.67,0.78,0.83, 0.89,0.94,1}, whereby first stage use half of the maximum informaiton and subsequential stages are equally spaced); then use this t vector to find non-binding WT upper bounds $\{u_1, \ldots, u_K\}$ by substiting $\{\frac{k}{K}\}$ in Equation 9.1 by t vector; then use it to to find error spent by $\beta(\frac{I_i}{I_K}, \delta_{min})= \beta(t_i, \delta_{min})= \beta(t_i)^{\gamma}$; then utilizing Equation 9.2 together with known upper bounds, we can get lower bound vector $\{l_1, \ldots l_K\}$; then we can search for a coefficient $R(K, \alpha, \beta)$ (Chapter 2, Jennison and Turnbull, 2000), which is the maximum information I_K divided by information needed for fixed sample desgin I_{fix} and hence get $\{I_1, \dots, I_K\}$. This is because: $R(K, \alpha, \beta)$ =I_K/I_{fix} , and $\{\frac{I_1}{I_K}, \frac{I_2}{I_K}, \dots, \frac{I_K}{I_K}\}$=$\{t_1, \dots, t_K\}$, so $\{I_1, \dots, I_K\}$= $R(K, \alpha, \beta)$*$\{t_1, \dots, t_K\}$. When upper, lower bounds and $\{t_1, \dots, t_K\}$ are given, $\{I_1, \dots, I_K\}$ is obtained by searching for $R(K, \alpha, \beta)$ to ensure power while also letting $l_K = u_K$ at final stage K to ensure only either rejecting or accepting null hypothesis at the final stage.

Section 4: Results

Based on the motivating example, we transform the trial objective of proving superiority of study drug relative to placebo to testing $H_0 = 0$ against $H_A = \delta_{min} = 3.24$. The required sample size for fixed design with α =0.025(one-sided) and β=0.1 is to accumulate 186 events. After obtaining the optimal shape parameters of ρ for Wang- Tsiatis upper bounds and γ for Kim-DeMets lower bounds satisfying all of 5 criteria for error rates overall and at stage one while minimizing ASN with respect to prior beliefs of δ (see Section 9.2.1 and 9.2.5 and Section 9.3), 3K parameters of upper, lower bounds and information vector can then be derived for this optimized group sequential design using optimal ρ and γ.

Group sequential tests allow stopping for efficacy and futility as early as stage one and then claim conclusion for hypothesis testing if bounds crossed at interim or otherwise continue up to the final stage. However, small numbers of patients accumulated at interims leave much to chance and greater uncertainty about the inferences. To avoid this not-large-enough sample size at interims causing more uncertainty issue, we coin our example with first interim occurred at the time when at least half of maximum information is used (i.e., t_1= $I_1/I_K = 0.5$). Furthermore, for simplicity, the remaining stages are equally spaced. For example, for K=10, we use standard information vector t^0={0.5, 0.55,0.61,0.67,0.78,0.83, 0.89,0.94,1} as the start point. t_1 can be increased to t_1^* to satisfy the power requirement of rejecting null under maximum effect size

(Equation 9.4), whereby the existence of t_1^* is proved in Appendix 9.1.

After obtaining 3K parameters of $\{l_1, \dots l_K\}$, $\{u_1, \dots, u_K\}$, $\{I_1, \dots, I_K\}$ for the optimized design sought-out by proposed algorithm (Figure 9.1), probability of stopping at stage i (i.e., $\Pr_\theta(T = i)$) can be calculated using sub-density at stage i (Pages 171-174, Jennison and Turnbull, 2000) and subsequently expected final information level, defined as $E_\theta\{I\} = \sum_{i=1}^{K} I_i * \Pr_\theta(T = i)$ summing over different stages can be obtained to evaluate efficiency of the proposed optimized design, where θ is at the scale of δ in a range that cover $\delta_{min}$and δ_{max}. In Figure 9.2, $E_\theta\{I\}/I_{fix}$, expected final information level divided by I_{fix} given θ (with x-axis ranging from -0.5*δ_{min} to 2*δ_{min}) is plotted against ratio of θ to δ_{min} for $\alpha_F = 0.2$ (solid line) and α_F =0.3 (dotted line) and for K=2,…,10, respectively. For an optimized group sequential design with K=2 and the probability of continuing to Stage Two under null being 0.2, when the parameter θ is the same as the minmial effect size (i.e. θ/δ_{min}=1), the expected final information level relative to that of fixed-sample design for the proposed optimized design is 0.786 (Figure 9.2a), which also means the expected number of events is 0.786*N_{fix}=0.786*186; and similarly for α_F=0.3, the expected information $E_\theta\{I\}/I_{fix}$=0.775. There is little interest for investigating θ less than δ_{min}, as we are not pursuing any investigational drug less than minimal effect size. Thus for θ ranging from δ_{min} to 1.5*δ_{min} is much of our interests. When we look at the effect size which is 1.5 times the minimal requirement (i.e., θ/δ_{min}=1.5), the expected final information level relative to that of fixed-sample design is 0.632 and 0.606 for α_F=0.2 and α_F=0.3, respectively (Figure 9.2a). This shows that designs with the same K, bigger effect size saves more resources; and for a given effect size, bigger α_F spent at first stage leads to smaller expected final information level.

One intuition is that designs with more interim analyses could result in smaller expected final information level $E_\theta\{I\}$. Surprisingly, we found this perception is only partially true. Table 9.2 lists $E_\theta\{I\}/I_{fix}$for effect size from as low as $\delta_{\min}$ up to 1.5 times of minimal effect size by increment of 0.05 in the ratio of $\theta/\delta_{\min}$. Let's take the extreme cases K=2 vs. K=10 in Table 2 to illustrate our points. Comparing K = 2 with K = 10 for $\alpha_F =$ 0.2, optimized group sequential tests with K = 10 consistently have lower expected final information level for $\theta/\delta_{\min}$ ranging from 1 to 1.35 (Table 9.2) than those of K = 2; however, the trend is reversed for ratios ranging from 1.40 to 2.0 (Note that data for ratios between 1.50 and 2.0 are not shown in Table 9.2). The same phenomenon is also observed for $\alpha_F = 0.3$. All in all, when ratio $\theta/\delta_{\min}$ is 1.50 and up, K = 10 has bigger expected final information level as compared with K = 2 while being smaller between ratios of 1.0 and 1.45 (shaded cells in Table 9.2). This actually shows that for a certain α_F, increasing the number of analyses can not save resources when the effect size is too big. Additionaly, the saving in sample size is very limited when K is greater or equal to 3 irrespective of effect size.

Table 35(Tab. 9.2): Efficiencies for optimal asymmetric optimal group designs

Table 9.2: $E_\theta\{I\}/I_{fix}$ for α_F =0.2 or 0.3 when $\theta/\delta_{\min}$ ranging from 1.0 to 1.5 with increments of 0.05

$\theta/\delta_{\min}$ =		1.0	1.05	1.1	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50
α_F= 0.2	K=2	0.786	0.766	0.745	0.726	0.707	0.690	0.675	0.662	0.650	0.640	0.632
	K=3	0.754	0.737	0.720	0.704	0.690	0.677	0.666	0.656	0.648	0.641	0.635
	K=4	0.745	0.729	0.713	0.699	0.686	0.674	0.664	0.656	0.648	0.642	0.637
	K=5	0.743	0.728	0.714	0.701	0.690	0.680	0.671	0.664	0.658	0.653	0.650
	K=6	0.738	0.723	0.709	0.696	0.684	0.674	0.665	0.657	0.651	0.646	0.642
	K=7	0.736	0.722	0.708	0.695	0.684	0.674	0.665	0.657	0.651	0.646	0.642
	K=8	0.736	0.722	0.709	0.697	0.686	0.677	0.669	0.662	0.657	0.653	0.649
	K=9	0.740	0.727	0.714	0.703	0.693	0.685	0.677	0.671	0.666	0.663	0.660
	K=10	0.734	0.719	0.706	0.694	0.683	0.673	0.665	0.658	0.652	0.647	0.643
α_F= 0.3	K=2	0.775	0.753	0.731	0.710	0.690	0.672	0.655	0.640	0.627	0.615	0.606
	K=3	0.739	0.720	0.701	0.684	0.668	0.654	0.641	0.629	0.620	0.611	0.605
	K=4	0.728	0.711	0.694	0.678	0.663	0.650	0.639	0.629	0.620	0.613	0.608
	K=5	0.723	0.707	0.690	0.675	0.661	0.649	0.638	0.629	0.622	0.615	0.610
	K=6	0.721	0.704	0.688	0.674	0.660	0.649	0.638	0.629	0.622	0.616	0.611
	K=7	0.719	0.703	0.687	0.672	0.659	0.647	0.637	0.628	0.621	0.615	0.610
	K=8	0.718	0.702	0.686	0.672	0.659	0.647	0.637	0.628	0.621	0.615	0.610

	K=9	0.717	0.701	0.686	0.671	0.658	0.647	0.637	0.629	0.621	0.615	0.611
	K=10	0.716	0.700	0.685	0.671	0.658	0.647	0.637	0.629	0.622	0.616	0.616

Back to Figure 9.2, which plots all scenarios on expected final information level relative to information of fixed-sample design for K = 2 up to 10 and α_F =0.2 or 0.3, except for θ/δ_{min} ranging from -0.5 to 0.7 in K = 2, the remainder of the design scenarios are uniformly most cost-effective (i.e., having smaller expected final information level) for $\alpha_F = 0.3$ than those of $\alpha_F = 0.2$. Looking at the shape of the curve in Figure 9.2 a-i, for each α_F, shapes of K $\geq$ 3 are all similar to each other and different from that of K = 2. So there are cost savings in terms of $E_\theta\{I\}$ from K = 2 to K = 3 for a given α_F, but there is no further savings in having a larger K when K $\geq$ 3. The range of $E_\theta\{I\}/I_{fix}$ for $\alpha_F =$ 0.2 is all smaller than that of $\alpha_F = 0.3$ showing a smaller variability in expected final information level when $\alpha_F = 0.2$. Irrespective of the value of α_F and K, maximum of $E_\theta\{I\}/I_{fix}$ occurs when $\theta/\delta_{min} = 0.6$. Except for K = 2, all maximum of $E_\theta\{I\}/I_{fix}$ is a little smaller for $\alpha_F = 0.3$ than that of $\alpha_F = 0.2$. Our examples confirmed that it is worthwhile to have K = 3 in order to reduce expected sample size but it seems not worthwhile to further increase it to K = 4, and similar phenomenon was also noticed in Anderson (2007).

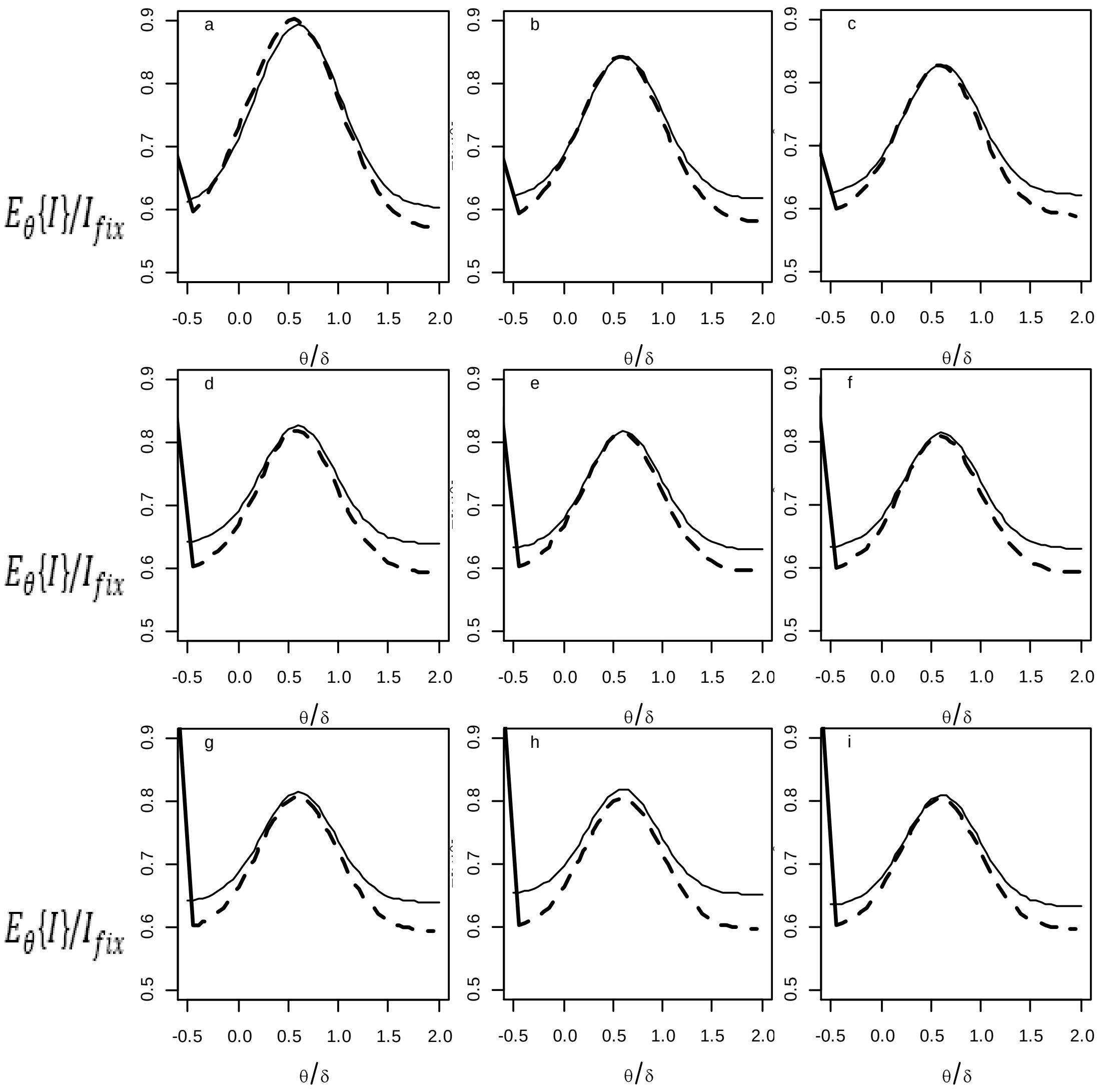


Figure 26(Fig. 9.2): Efficiencies of optimized asymmetric group sequential designs

Figure 9.2: $E_\theta\{I\}/I_{fix}$ vs. θ/δ_{min} for optimized asymmetric group sequential designs minimizing ASN when α =0.025(one-sided), $\alpha_F = 0.2$(solid line), 0.3(dotted line), β=0.1, K=2,3,4,5,6,7,8,9,10, and $I_1/I_K = 0.5$ and the remaining stages equally spaced. Note: a: K=2, b: K=3, c: K=4, d: K=5, e: K=6, f: K=7, g: K=8, h: K=9, i: K=10.

Operating characteristics for scenarios in Figure 9.2 and Table 9.2 are depicted in detail in Tables 9.3 and 9.4 accompanying with 3K parameters of lower boundaries, upper boundaries and information vector, and probability of rejecting null under maximum effect size at Stage One to control probability of continuing to stage two when null hypothesis is true (α_F= 0.3, 0.2). As α_F= 0.3, a more lenient probability of continuing to stage two under the null, was advocated by

Liu, et, al (2012), we start our discussion with α_F= 0.3 (Tables 9.3). With continuing probability at stage one when null is true controlled at 0.3 level and overall power equal to 0.9, maximum information relative to fixed-sample design, I_{max}/I_{fix} , is 1.135, 1.153, 1.174 and 1.183 for K = 2,3,4,5, respectively in our method while Anderson (2007) had 1.106, 1.180, 1.218, and 1.237. Our method only has a slightly bigger I_{max}/I_{fix} than that of Anderson (2007) at K = 2 while the remaining Ks being smaller than Anderson (2007), showing advantage of our optimized group sequential tests in terms of reducing maximum information level with respect to prior beliefs of effect size. The real problem for Anderson (2007) is their lower information level at stage one, only with 0.553, 0.393, 0.305 and 0.247 for I_1/I_{fix} for K = 2, 3, 4 and 5, respectively, while we have at 0.567 for K = 2 and this value increases to 0.595 when K = 10. Decisions made only using 0.247 percent of total information for fixed sample design will leave any decision on this in doubt, especially significance in efficacy, more to chance rather than real drug effect; and this shortcoming for Anderson (2007) is the primary propulsion for us to develop a better optimized design here. The maximum information, even not fixed in advance, turns out to be well-controlled using our searching method for optimal shapes for ρ and γ (Figure 9.1). For example, it is only 1.19 even for K = 10 and power = 0.9 (Table 9.3).

Due to implementing of non-binding upper boundary, overall type I error, as expected, is a little less than pre-specified 0.025 level irrespective of power = 0.8 or 0.9 and α_F = 0.3 or 0.2 (Tables 9.3 and 9.4). Comparing with α_F= 0.3, α_F= 0.2 has bigger I_{max} for any combination of K and power (I_{max}/I_{fix} =1.204, 1.264 for K = 2 and 10, respectively). The first stage lower bound, l_1 is higher in α_F = 0.2 than that of α_F = 0.3 to limit the chance of going to stage two under null (l_1= 0.842 for α_F = 0.2 and l_1= 0.524 for α_F= 0.3). As expected, the maximum information is lower in power of 0.8 than that of power = 0.9. One surprising finding in Tables 9.3 and 9.4 is for power equal to 0.8: all I_{max}/I_{fix} are less than 1 for all combinations of K and α_F.

Table 36(Tab. 9.3): Optimized asymmetric groups sequential designs minimizing ASN

Table 9.3: Optimized asymmetric groups sequential designs minimizing ASN with $\alpha = 0.025$(one-sided), $\beta = 0.1$, or 0.2, k=2,3,4,5,6,7,8,9, 10, powered at δ_{min}=3.24 and t_1=0.5 and the remaining stages are equally spaced.

	$\beta = 0.1$ (Power=0.9)						$\beta = 0.2$ (Power=0.8)				
K=2	$\rho = 0.5047$, $\gamma = 1.862$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.956 and δ_{Max}=5.147						$\rho = 0.4516$, $\gamma = 1.805$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.869 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0148	0.606	0.524	2.175	0.567		0.0133	0.454	0.524	2.217	0.420
2	0.0244	0.900	2.182	2.182	1.135		0.0243	0.800	2.144	2.144	0.841
K=3	$\rho = 0.4391$, $\gamma = 1.945$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.947 and δ_{Max}=5.147						$\rho = 0.4116$, $\gamma = 1.889$, α_F=0.3, $P_{\delta_{Max}}(Z_1 \geq b_1)$= 0.854 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0107	0.566	0.524	2.301	0.579		0.0100	0.423	0.524	2.327	0.432
2	0.0183	0.799	1.359	2.245	0.868		0.0178	0.665	1.338	2.245	0.648
3	0.0238	0.900	2.206	2.206	1.158		0.0237	0.800	2.188	2.188	0.864
K=4	$\rho = 0.4346$, $\gamma = 2.003$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.945 and δ_{Max}=5.147						$\rho = 0.3959$, $\gamma = 1.926$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.846 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0095	0.556	0.524	2.346	0.587		0.0086	0.406	0.524	2.382	0.437
2	0.0154	0.737	1.063	2.302	0.783		0.0145	0.591	1.046	2.312	0.583
3	0.0203	0.847	1.613	2.268	0.979		0.0198	0.724	1.585	2.259	0.729
4	0.0236	0.900	2.242	2.242	1.174		0.0236	0.800	2.216	2.216	0.874
K=5	$\rho = 0.4343$, $\gamma = 2.036$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.944 and δ_{max}=5.147						$\rho = 0.3910$, $\gamma = 1.949$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.842 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0088	0.549	0.524	2.372	0.592		0.0079	0.397	0.524	2.414	0.440
2	0.0137	0.699	0.905	2.338	0.740		0.0127	0.547	0.891	2.356	0.550
3	0.0178	0.801	1.325	2.310	0.888		0.0170	0.664	1.300	2.309	0.660
4	0.0213	0.866	1.746	2.286	1.036		0.0208	0.750	1.713	2.271	0.770
5	0.0236	0.900	2.267	2.267	1.183		0.0235	0.800	2.238	2.238	0.880
K=6	$\rho = 0.4247$, $\gamma = 2.049$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.941 and δ_{max}=5.147						$\rho = 0.3860$, $\gamma = 1.962$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.837and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0082	0.539	0.524	2.400	0.594		0.0074	0.389	0.524	2.438	0.442
2	0.0123	0.670	0.805	2.367	0.712		0.0115	0.517	0.795	2.388	0.530
3	0.0159	0.763	1.144	2.340	0.831		0.0151	0.620	1.124	2.346	0.619
4	0.0191	0.830	1.478	2.317	0.950		0.0185	0.702	1.452	2.311	0.707
5	0.0217	0.876	1.824	2.296	1.069		0.0214	0.763	1.793	2.280	0.796
6	0.0235	0.900	2.278	2.278	1.187		0.0234	0.800	2.253	2.253	0.884
K=7	$\rho = 0.3896$, $\gamma = 2.040$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.941 and δ_{max}=5.147						$\rho = 0.3828$, $\gamma = 1.971$, α_F=0.3, $P_{\delta_{max}}(Z_1 \geq b_1)$= 0.834and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	b_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	b_i	I_i/I_{fix}
1	0.0072	0.519	0.524	2.449	0.592		0.0070	0.384	0.524	2.456	0.443
2	0.0107	0.638	0.734	2.408	0.691		0.0106	0.496	0.729	2.412	0.517
3	0.0139	0.727	1.015	2.372	0.790		0.0138	0.587	1.004	2.374	0.591
4	0.0168	0.796	1.295	2.342	0.888		0.0167	0.663	1.278	2.342	0.665
5	0.0195	0.846	1.573	2.315	0.987		0.0194	0.725	1.554	2.313	0.739
6	0.0218	0.881	1.867	2.290	1.086		0.0218	0.772	1.850	2.287	0.813
7	0.0234	0.900	2.269	2.269	1.185		0.0234	0.800	2.264	2.264	0.887
K=8	$\rho = 0.3856$, $\gamma = 2.048$, α_F=0.3,						$\rho = 0.3800$, $\gamma = 1.9773$, α_F=0.3,				

	$\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.933 and δ_{max}=5.147						$\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.831 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0069	0.513	0.524	2.465	0.593		0.0067	0.379	0.524	2.470	0.444
2	0.0100	0.621	0.685	2.427	0.678		0.0099	0.479	0.681	2.431	0.508
3	0.0129	0.703	0.925	2.395	0.763		0.0127	0.561	0.915	2.397	0.571
4	0.0155	0.768	1.167	2.366	0.848		0.0154	0.632	1.152	2.367	0.635
5	0.0179	0.819	1.405	2.341	0.933		0.0178	0.691	1.388	2.340	0.698
6	0.0202	0.858	1.647	2.317	1.017		0.0201	0.740	1.628	2.316	0.761
7	0.0221	0.885	1.909	2.296	1.102		0.0221	0.777	1.892	2.294	0.825
8	0.0234	0.900	2.277	2.277	1.187		0.0234	0.800	2.273	2.273	0.888
K=9	ρ = 0.3833, γ = 2.054, α_F=0.3, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.932 and δ_{max}=5.147						ρ = 0.3768, γ = 1.9816, α_F=0.3, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.829 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0066	0.509	0.524	2.477	0.594		0.0065	0.374	0.524	2.483	0.445
2	0.0095	0.608	0.648	2.443	0.669		0.0094	0.465	0.646	2.448	0.500
3	0.0121	0.683	0.856	2.413	0.743		0.0119	0.541	0.848	2.416	0.556
4	0.0144	0.745	1.069	2.386	0.817		0.0143	0.606	1.056	2.388	0.612
5	0.0167	0.795	1.279	2.362	0.892		0.0165	0.662	1.263	2.362	0.667
6	0.0188	0.835	1.489	2.340	0.966		0.0187	0.711	1.470	2.339	0.723
7	0.0207	0.866	1.704	2.320	1.040		0.0206	0.750	1.685	2.318	0.778
8	0.0223	0.888	1.941	2.302	1.114		0.0223	0.781	1.925	2.298	0.834
9	0.0233	0.900	2.284	2.284	1.189		0.0234	0.800	2.280	2.280	0.890
K=10	ρ = 0.3814, γ = 2.059, α_F=0.3, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.931 and δ_{max}=5.147						ρ = 0.3741, γ = 1.9849, α_F=0.3, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.826 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0064	0.506	0.524	2.487	0.595		0.0063	0.371	0.524	2.494	0.445
2	0.0091	0.597	0.620	2.456	0.661		0.0089	0.454	0.618	2.462	0.495
3	0.0114	0.667	0.802	2.428	0.727		0.0113	0.524	0.795	2.432	0.544
4	0.0136	0.725	0.992	2.404	0.793		0.0134	0.584	0.980	2.406	0.594
5	0.0156	0.774	1.180	2.381	0.860		0.0155	0.638	1.164	2.382	0.643
6	0.0176	0.814	1.366	2.360	0.926		0.0174	0.685	1.348	2.360	0.693
7	0.0194	0.846	1.554	2.341	0.992		0.0193	0.725	1.535	2.339	0.742
8	0.0210	0.871	1.749	2.323	1.058		0.0210	0.758	1.730	2.320	0.792
9	0.0224	0.890	1.968	2.306	1.124		0.0224	0.784	1.951	2.303	0.841
10	0.0233	0.900	2.291	2.291	1.190		0.0234	0.800	2.286	2.286	0.890

Table 37(Tab. 9.4): Optimized asymmetric groups sequential designs minimizing ASN

Table 9.4: Optimized asymmetric groups sequential designs minimizing ASN with $\alpha = 0.025$(one-sided), $\beta = 0.1, or\ 0.2$, k=2,3,4,5,6,7,8,9, 10, powered at δ_{min}=3.24 and t_1=0.5 and the remaining stages are equally spaced.

	$\beta = 0.1$ (Power=0.9)						$\beta = 0.2$ (Power=0.8)				
K=2	ρ = 0.5546, γ = 1.091, α_F=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$=0.968 and δ_{max}=5.147						ρ = 0.4881 , γ = 1.099, α_F=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.893 and δ_{max}=				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	a_i	u_i	I_i/I_{fix}
1	0.0162	0.647	0.842	2.140	0.602		0.0144	0.490	0.842	2.187	0.444
2	0.0238	0.900	2.222	2.222	1.204		0.0235	0.800	2.169	2.169	0.889
K=3	ρ =,0.4909 γ = 1.1752, α_F=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$=0.963 and δ_{max}=5.147						ρ = 0.4709, γ = 1.1953, α_F=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.888 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0120	0.613	0.842	2.257	0.616		0.0115	0.470	0.842	2.274	0.460
2	0.0190	0.822	1.494	2.249	0.923		0.0187	0.694	1.481	2.247	0.689
3	0.0231	0.900	2.243	2.243	1.231		0.0230	0.800	2.228	2.228	0.919

K=4	$\rho =\ 0.4627,\ \gamma = 1.211,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.959 and δ_{max}=5.147						$\rho =\ 0.4377\ ,\ \gamma =\ 1.219,\ \alpha_{F1}$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.877 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0102	0.593	0.842	2.320	0.621		0.0096	0.446	0.842	2.343	0.463
2	0.0159	0.766	1.230	2.296	0.828		0.0154	0.625	1.223	2.301	0.618
3	0.0203	0.860	1.697	2.277	1.035		0.0199	0.742	1.678	2.269	0.772
4	0.0228	0.900	2.261	2.261	1.242		0.0227	0.800	2.244	2.244	0.927
K=5	$\rho = 0.5835,\ \gamma = 1.3193,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.969 and δ_{max}=5.147						$\rho =\ 0.4553,\ \gamma =\ 1.249\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.879 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0122	0.634	0.842	2.250	0.638		0.0093	0.447	0.842	2.353	0.468
2	0.0169	0.756	1.115	2.293	0.798		0.0141	0.590	1.093	2.330	0.585
3	0.0201	0.833	1.480	2.328	0.958		0.0180	0.694	1.432	2.311	0.702
4	0.0223	0.878	1.871	2.358	1.117		0.0210	0.764	1.795	2.295	0.819
5	0.0234	0.900	2.384	2.384	1.277		0.0228	0.800	2.281	2.281	0.936
K=6	$\rho = 0.4660\ ,\ \gamma = 1.2579\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.957 and δ_{max}=5.147						$\rho =\ 0.4278,\ \gamma = 1.249\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.870 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0091	0.583	0.842	2.362	0.629		0.0083	0.429	0.842	2.397	0.468
2	0.0133	0.706	1.011	2.348	0.754		0.0124	0.555	1.008	2.366	0.562
3	0.0167	0.791	1.287	2.336	0.880		0.0159	0.652	1.275	2.340	0.656
4	0.0194	0.848	1.574	2.325	1.006		0.0189	0.724	1.552	2.317	0.749
5	0.0215	0.883	1.884	2.316	1.132		0.0212	0.774	1.855	2.298	0.843
6	0.0228	0.900	2.307	2.307	1.257		0.0226	0.800	2.280	2.280	0.937
K=7	$\rho = 0.4609,\ \gamma = 1.267\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.956 and δ_{max}=5.147						$\rho =\ 0.4237,\ \gamma =\ 1.256\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.867 band δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0086	0.577	0.842	2.381	0.630		0.0079	0.423	0.842	2.416	0.469
2	0.0123	0.687	0.954	2.367	0.735		0.0115	0.534	0.953	2.387	0.548
3	0.0154	0.765	1.180	2.355	0.840		0.0146	0.621	1.171	2.363	0.626
4	0.0180	0.822	1.417	2.344	0.945		0.0173	0.690	1.400	2.342	0.704
5	0.0201	0.862	1.663	2.334	1.050		0.0196	0.743	1.638	2.323	0.782
6	0.0218	0.887	1.934	2.326	1.155		0.0214	0.780	1.903	2.306	0.861
7	0.0228	0.900	2.318	2.318	1.260		0.0226	0.800	2.291	2.291	0.939
K=8	$\rho = 0.5405,\ \gamma =\ 1.323\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.963 and δ_{max}=5.147						$\rho =\ 0.6103,\ \gamma = 1.359\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.905 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0100	0.605	0.842	2.326	0.639		0.0115	0.495	0.842	2.274	0.486
2	0.0135	0.696	0.922	2.339	0.730		0.0151	0.580	0.931	2.308	0.555
3	0.0162	0.761	1.117	2.350	0.822		0.0176	0.643	1.125	2.338	0.625
4	0.0184	0.810	1.325	2.360	0.913		0.0196	0.693	1.333	2.365	0.694
5	0.0201	0.846	1.539	2.369	1.004		0.0211	0.733	1.550	2.390	0.764
6	0.0215	0.873	1.765	2.377	1.095		0.0222	0.764	1.783	2.414	0.833
7	0.0226	0.891	2.020	2.385	1.187		0.0231	0.786	2.051	2.435	0.902
8	0.0232	0.900	2.392	2.392	1.278		0.0236	0.800	2.455	2.455	0.971
K=9	$\rho =\ 0.6386,\ \gamma = 1.4012\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.971 and δ_{max}=5.147						$\rho =\ 0.4442\ ,\ \gamma =\ 1.276\ ,\ \alpha_F$=0.2, $\mathrm{P}_{\delta_{max}}(Z_1 \geq b_1)$= 0.869 and δ_{max}=5.147				
	α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1-\beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0118	0.638	0.842	2.264	0.651		0.0078	0.425	0.842	2.418	0.473
2	0.0152	0.713	0.902	2.301	0.733		0.0109	0.514	0.887	2.402	0.532
3	0.0175	0.766	1.075	2.335	0.814		0.0134	0.585	1.043	2.388	0.591
4	0.0193	0.806	1.263	2.366	0.896		0.0157	0.644	1.213	2.375	0.650
5	0.0207	0.838	1.456	2.395	0.977		0.0176	0.693	1.388	2.364	0.709
6	0.0219	0.862	1.658	2.422	1.058		0.0194	0.733	1.568	2.353	0.768
7	0.0227	0.880	1.874	2.447	1.140		0.0208	0.765	1.761	2.343	0.827

8	0.0233	0.893	2.123	2.470	1.221		0.0220	0.787	1.984	2.334	0.886
9	0.0237	0.900	2.492	2.492	1.303		0.0227	0.800	2.326	2.326	0.945
K=10	$\rho = 0.4456$, $\gamma = 1.278$, α_F=0.2, P$_{\delta_{max}}(Z_1 \geq b_1)$= 0.952 and δ_{max}=5.147						$\rho = 0.4305$, $\gamma = 1.274$, α_F=0.2, P$_{\delta_{max}}(Z_1 \geq b_1)$= 0.864 and δ_{max}=5.147				
	α_i	$1 - \beta_i$	l_i	u_i	I_i/I_{fix}		α_i	$1 - \beta_i$	l_i	u_i	I_i/I_{fix}
1	0.0077	0.561	0.842	2.424	0.632		0.0074	0.417	0.842	2.439	0.472
2	0.0105	0.646	0.859	2.411	0.702		0.0102	0.499	0.863	2.421	0.525
3	0.0129	0.710	0.998	2.398	0.772		0.0125	0.566	0.997	2.405	0.577
4	0.0149	0.762	1.152	2.387	0.843		0.0146	0.622	1.147	2.390	0.630
5	0.0168	0.803	1.310	2.376	0.913		0.0165	0.670	1.300	2.377	0.682
6	0.0185	0.836	1.472	2.367	0.983		0.0182	0.711	1.458	2.365	0.735
7	0.0199	0.862	1.638	2.358	1.053		0.0197	0.744	1.621	2.354	0.787
8	0.0211	0.880	1.816	2.350	1.123		0.0210	0.770	1.797	2.343	0.839
9	0.0221	0.893	2.022	2.342	1.194		0.0220	0.789	2.002	2.333	0.892
10	0.0227	0.900	2.335	2.335	1.264		0.0226	0.800	2.324	2.324	0.944

Section 9.5: Discussion

Maximum sample size in our method is not fixed as Barber and Jennison (2002), Jennison (1987), Eales and Jennison (1992) and Jennison and Turnbull (2004) have done. And the maximum sample size is determined by optimization with help of shape parameters after implementing the iterative algorithm in Figure 9.1, which turns out to be better than Anderson (2007) (Tables 9.3 and 9.4) in terms of reducing resources in addition to more constraints on stage one probabilities. Wang and Tsiatis (1987) and Kim and DeMets (1987) are used here and there does not appear to be a need in using a more complex spending function family as in Jennison (1987). There are better features in our method as compared with previous ones mentioned above (Barber and Jennison (2002), Jennison (1987), Jennison (1992), Jennison and Turnbull (2004) and Anderson (2007)): power of rejecting at stage one is ensured when maximum effect size is true; error of continuing the trial when no drug effect exists is well-controlled at stage one; and non-binding efficacy boundaries are used to account for overrunning data that normally occur in every real trial. In evaluating the number of analyses to perform, there is a benefit to increase analyses from two stages to three stages and perhaps little benefit in having more than 3 stages in most cases, while Anderson’s method (2007) shows no benefit in having more than 4 stages. Fewer interim analyses should save a lot of on human resources and needed-time in conducting additional interim data cleaning and analysis. However, we have not done any example with unequal spacing between Stage 2 and the maximum stage. Though it is very easy to find optimized group sequential design using our method if unequal spacing is desirable for some operational reasons, Barber and Jennison (2002) noted that optimal designs allowing unequal spacing provide minimal advantage over equal spacing. R codes are available for the first author per your requests.

References

Anderson, K. (2007). Optimal spending functions for asymmetrical group sequential designs. *Biometrical Journal* 49, 337-45.

Armitage, P., Mcpherson, C.K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Statist. Soc. A.* 132, 235-44.

Barber, S. and Jennison, C (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika* 89, 49-60.

Eales, J. D. and Jennison, C. (1992). An improve method for deriving one-sided group sequential designs. *Biomterika* 79, 13-24.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* 44, 793-97.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton, FL.

Jennison, C. and Turnbull, B.W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Stat Med.* 25, 917-32.

Kim, K. and Demets, D.L. (1987). Design and Analysis of group sequential tests based on Type I error spending rate functions. *Biometrika* 74, 149-54

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70, 659-64.

Liu, Q. and Anderson, K. M. (2008a). On adaptive extensions of group sequential trials for clinical investigations. *J. Am. Statist. Assoc*. 103, 1621-630.

Liu, Q. and Anderson, K. M. (2008b). Theory of inference for adaptively extended group sequential designs with applications in clinical trials. *J. Am. Statist. Assoc*. Supplemental Archive; http://pubs.amstat.org/toc/jasa/103/484

Liu, Q. and Chi, Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* 57, 172-77.

O’Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 549-56.

Pocock, S. J. (1977). Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 38, 153-162.

Wang, S.K. and Tsiatis A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193-199.

Appendix 9.1

Claim $\exists\ t_1^*,\ \ t_1^0 < t_1^* \le 1$, such that $P_{\delta_{max}}(Z_1 \ge u_1) \ge 1 - \beta$
Let's prove one preliminary result first: Given $\delta_{Max} > \delta_{min} > 0,\ \ P_{\delta_{max}}(Z_1 \ge u_1) > P_{\delta_{min}}(Z_1 \ge u_1)$
Proof: this is because $Z_1 \sim N(\sqrt{I_1}\delta, 1)$, which results in $P_{\delta_{max}}(Z_1 \ge u_1)$=1-$\Phi(u_1 - \delta_{max}\sqrt{I_1})$ and $P_{\delta_{min}}(Z_1 \ge u_1)$=1-$\Phi(u_1 - \delta_{min}\sqrt{I_1})$, then directly we have $P_{\delta_{max}}(Z_1 \ge u_1) > P_{\delta_{min}}(Z_1 \ge u_1)$ because of $\delta_{max} > \delta_{min} > 0$. Similarly, we have $P_{\delta_{max}}(\ \cup_{i=1}^{K} Z_i \ge u_i) > P_{\delta_{min}}(\ \cup_{i=1}^{K} Z_i \ge u_i)$. Per optimization algorithm in Figure 1, $P_{\delta_{min}}(\ \cup_{i=1}^{K} Z_i \ge u_i) = 1 - \beta$. Let $P_{\delta_{max}}(\ \cup_{i=1}^{K} Z_i \ge u_i)$=1-$\beta'$, where $\beta' < \beta$ to satisfy 1-$\beta' > 1 - \beta$. Let $1 - \beta' = 1 - \beta + \Delta$, where difference $\Delta = (1 - \beta') - (1 - \beta)$>0.
Because $P_{\delta_{max}}(\ \cup_{i=1}^{K} Z_i \ge u_i) = P_{\delta_{max}}(Z_1 \ge u_1) + P_{\delta_{max}}\left(\ \cup_{i=1}^{K-1} Z_i \ge u_i\right)$ = A+B if $P_{\delta_{max}}(Z_1 \ge u_1)$ = A and $P_{\delta_{max}}\left(\ \cup_{i=1}^{K-1} Z_i \ge u_i\right)$= B respectively. ∴ $P_{\delta_{Max}}(Z_1 \ge u_1)$ = A = $1 - \beta + \Delta$ −B>0. Our objective becomes to prove: $\exists\ t_1^*,\ \ t_1^0 < t_1^* \le 1$, such that $A \ge 1 - \beta$. There are two cases for this. Case One: If using t_1^0, initial (least) standard fraction of information used at stage one, we already have $P_{\delta_{max}}(Z_1 \ge u_1) \ge 1 - \beta$, then there is nothing to prove. We just use t_1^0 together with the chosen way of partition for the remaining information to search for each optimized design. Case Two: at t_1^0, we have $P_{\delta_{max}}(Z_1 \ge u_1)$ = A<$1 - \beta$, then we have to show that when we increase t_1^0 to t_1^*, we can have $P_{\delta_{max}}(Z_1 \ge u_1)$ = A$\ge 1 - \beta$.
To prove Case Two, we know that $I_1 = I_{max} * t_1$, where I_{max} is determined by α, β and α_F and has nothing to do with δ_{max} (Figure 9.1). So, again, for $P_{\delta_{max}}(Z_1 \ge u_1)$ = A = 1-$\Phi(u_1 - \delta_{Max}\sqrt{I_1})$= 1-$\Phi(u_1 - \delta_{Max}\sqrt{I_{max} * t_1})$. Given I_{max}, A increases as t_1 increases. At the extremity, $t_1 = 1$, a case that group sequential design degenerates to the usual fixed sample design, $P_{\delta_{max}}(Z_1 \ge u_1) > P_{\delta_{min}}(Z_1 \ge u_1)$ = 1-β, which is what we proved above in the preliminary. For any $t_1$in between, that is $t_1^0 < t_1 \le 1$,
We have a continuous probability function A, which is a function in t_1 , in a closed interval $[\ t_1^0, 1]$, A has a real value at t_1^0 less than $1 - \beta$, on the other hand has a real value at t=1 greater or equal to $1 - \beta$. Per Intermediate Value Theroem from Real Analysis, we can conclude that there is a t_1^* , with $t_1^0 < t_1^* \le 1$, such that $P_{\delta_{max}}(Z_1 \ge u_1) = 1 - \beta$ is exactly achieved at t_1^*. When $t_1 > t_1^*$, A = $P_{\delta_{max}}(Z_1 \ge u_1) > 1 - \beta$.

Chapter 10
A Two-stage Adaptive Design with a New Combination Test

(to be submitted)

Abstract: Inspired by Bauer and Kohne (1994), a method applying Fisher's combination rule to form a two-stage adaptive procedure (BK method), utilizing Box and Muller (1958), one of the most popular methods of generating standard normal random variable using two independent uniform (0, 1) deviates, a new method is proposed here to combine two p-values from two disjoint samples for designing a trial with two stages. Procedure is defined with carefully consideration of controlling overall type I error rate under null hypothesis. Operational characteristics including power and expected sample size under both null and alternative hypotheses were investigated. Simulations were used to confirm type I error control. Comparisons of new combination method with BK method were also investigated.

Key Words: Two-stage Adaptive Design; Combination Test; Sample Size Re-estimation.
Subject classification codes: 05B99 62E20

Section 10.1: Introduction

In adaptive or flexible designs, study is monitored at interim while data are still being accrued and the study design, such as sample size, allocation of treatment et.al, can be modified accordingly to new internal/external information after the interim analysis. Statistical approaches must be shown to maintain the integrity of the trial such as controlling type I error as well as gaining adequate power. Among many publications, there are three methods wildly discussed and cited in the literature to deal with adaptations: Conditional power approach by Proshan and Hunsberger (1995); and two for combination tests: i) Fisher's combination rule by Bauer and Kohne (1994) and ii) the inverse normal method by Lehmacher and Wassmer (1999). In Proshan and Hunsberger (1995), the circular conditional error function, which increases for the increasing value of test statistic at stage 1, was defined for p-value of p_2 . Null hypothesis would be rejected if p_1 was less than or equal to α_1(alpha spent at stage 1) or p_2 was less than or equal to the conditional error at stage 2. Bauer and Kohne (1994) made use of the fact that $-2\log(p_i), i = 1,2$ has a Chi-squared distribution with 2 degree of freedom. Thus the product of p_1 and p_2 from disjoint data from stage 1 and stage 2 respectively was with a Chi-squared distribution with degree of freedom 4. To control the overall alpha level, a combination test $(p_1, p_2) = \mathrm{p}_1 * \mathrm{p}_2 \leq \exp(-0.5\chi_4^2(1-\alpha))$ could be utilized, where $\chi_4^2(1-\alpha)$ is the 100*(1- $\alpha)th$ percentile of the Chi-squared distribution with 4 degree of freedom. Inverse normal method by Lehmacher and Wassmer (1999) was proposed under group sequential setting. It is simply the weighted-z test to replace original test,$C(p_1, p_2) = \sqrt{\mathrm{w}_1}\mathrm{Z}_1 + \sqrt{1-\mathrm{w}_1}\mathrm{Z}_2$, with which $\mathrm{Z}_\mathrm{i} = \Phi(1-\mathrm{p}_\mathrm{i})$ (i.e., the inverse of standard normal cumulative distribution function) and w_1 is pre-fixed weight for stage 1 data. Under null hypothesis and the predefined weight w_1, $\sqrt{\mathrm{w}_1}\mathrm{Z}_1 + \sqrt{1-\mathrm{w}_1}\mathrm{Z}_2$ would be a standard normal. Even though sample size update using interim results seemed creating dependence between two statistics between stages, the inverse normal of $1-\mathrm{p}_\mathrm{i}$ value always derived a standard normal variable to ensure inter-stage independence in testing statistic.

Similar to combining independent p-values using Fisher's combination test, our method utilized Box and Muller (1958) (BM transformation) to combine two p-values. Section 10.2 stated the formulation of this two-stage procedure. Starting from objective of the test, given overall alpha level and stage one futility boundary, alpha-spent at stage 1 will be derived. Section 10.3 illustrated how the power and expected sample size could be calculated under null and alternative hypothesis respectively. Examples of calculating operation characteristics were followed in Section 10.4. And simulations were used to confirm that type I error is controlled as desired. Discussion in Section 10.5 concluded this paper.

Section 10.2: Formulation

Considering the situation to compare mean μ_1 of treatment group with mean μ_2 of control group with a known common variance of σ^2, a two-stage test procedure for the one-sided testing of superiority of treatment over control (positive difference means better) is structured with hypotheses: $\mathrm{H_0}: \mu_1\text{-}\mu_2 = 0$ versus $\mathrm{H_A}: \mu_1\text{-}\mu_2 = 0$

The standardized effect size will be $\delta = \frac{\mu_1-\mu_2}{\sigma}$. Because each pair of subjects is identical and independently (i.i.d.) distributed with normal mean $\mu_1 - \mu_2$ and correspondingly variance of $2\sigma^2$, with n_1 subjects accumulated at interim, the test statistic is defined as $T_1 = \frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{2\sigma^2/n_1}}$, which should be Normal($\sqrt{n_1/2}\ \delta$, 1) and p-value for this test as p_1 =1-$\Phi(\frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{2\sigma^2/n_1}})$. Similar definitions are defined for T_2 and p_2. Under null hypothesis, p-values under null hypothesis are known to be uniformly distributed from 0 to 1.

Assuming p_1 and p_2 are independent, for example case 1)deriving from two different cohorts of subjects as in current formulation 2)don't come from two different cohorts of subjects but are indeed independent asymptotically as the formulation for survival analysis. Here we propose a new way to combine two-stage data so that adaptation can be implemented after interim analysis to account for updated information from interim results or from external information. This is based on the fact of $C(p_1, p_2) = X_c = \sqrt{-2\log(p_2)}\cos 2\pi\, p_1$, where $p_1 \perp p_2$ ("$\perp$" indicating independence) and X_c is distributed as a standard normal variable under null hypothesis with subscript c indicating 'combined' and X_c itself denoting the combined test statistic at the end of stage 2. At the end of Stage 2, null hypothesis H_0 will be rejected if $\sqrt{-2\log(p_2)}\cos 2\pi\, p_1 \leq z_{1-\alpha}$, with $z_{1-\alpha}$ denoting the 100*(1- α)th percentile of the standard normal distribution. Or null hypothesis will get rejected at first stage if $p_1 \leq \alpha_1$ (with $\alpha_1 < \alpha$) if early rejection is planned ahead. Let α_1 be the alpha-spent at interim and α be the overall alpha level for both stages. If stopping for futility is also planned at interim with $p_1 \geq \alpha_0$, given a value of α_0 that provides a lower bound for p_1 to stop the trial with the larger value of p_1 indicating acceptance of H_0 at interim, the two-stage procedure can be summarized as the follows:

If $p_1 \geq \alpha_0$, the trial stops with acceptance of H_0,

If $p_1 \leq \alpha_1 (\alpha_1 < \alpha)$, the trial stops with rejection of H_0,

Otherwise, $p_1 < \alpha_1 \leq \alpha_0$, the second stage procedure can be performed; and in the second stage, H_0 can be rejected if $p_2 \leq \exp[-0.5 * (\frac{z_{1-\alpha}}{\cos 2\pi p_1})^2]$.

So, to get an overall level- α test, the value of α_1 has to be determined such that

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_{0}^{\exp[-0.5*(\frac{z_{1-\alpha}}{\cos 2\pi p_1})^2]} dp_2 dp_1 = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \exp[-0.5 * (\frac{z_{1-\alpha}}{\cos 2\pi\, p_1})^2] dp_1 = \alpha \quad (10.1)$$

If α_1 is given, α_0 can be determined using bisection searching together through above equation. Also, from above deduction, conditional error left for Stage 2 after observing p_1 is $A(p_1) = \exp[-0.5 * (\frac{z_{1-\alpha}}{\cos 2\pi p_1})^2]$, a function of p_1 and $z_{1-\alpha}$, is the critical value to be compared the combination test $C(p_1, p_2)$ combining $p_1$and p_2 (i.e., $C(p_1, p_2) = \sqrt{-2\log(p_2)} \cos 2\pi\, p_1$). Type I error will be well-controlled as long as $p_2 \leq A(p_1)$, even after n_2, sample size for Stage 2, is adapted to n_2^* based on interim results. Note that because α_1, type I error spent at stage 1, is normally less than or equal to 0.1, it can be seen that in the range of $0 \leq \alpha_1 \leq 0.1$, the conditional error for Stage 2 decreases for the increasing p_1. This shows the validity of proposed combination method, in which that a bigger p_1 at Stage 1 showing less evidence of treatment effect, rejection of H_0 at Stage 2 will become harder.

To interpret newly proposed BM method better, taking first row in Table 10.1 as an example, null hypothesis will be rejected at stage 1 if $p_1 \leq 0.0335$, or be accepted if $p_1 \geq 0.30$ or $t_1 \geq z_{\alpha_1}$; or go to gather Stage 10.2 data if $0.0335 < p_1 < 0.30$. At the end of Stage 10.2, data gathered from Stage 10.2 only will be used to obtain p_2, and null will be rejected if $p_2 \leq$

$\exp[-0.5 * (\frac{z_{1-\alpha}}{\cos 2\pi p_1})^2] = A(p_1)$; or equivalently the combined test statistic $c(p_1, p_2) = X_c = \sqrt{-2\log(p_2)} \cos 2\pi\, p_1 \leq z_{1-\alpha} = 1.644854$ in combining data in a way through p_1 and p_2; and will fail to reject null otherwise.

When α_0 is given, α_1 can be obtained using integration and bisection root searching using Equation 10.1. Given that $\alpha = 0.05$, for $\alpha_0 = 0.30, 0.35, 0.40, 0.45, 0.50$, respectively, one can find corresponding α_1 be 0.0335, 0.0332, 0.0286, 0.0166 and 0.0001 (Table 10.1). It is very interesting to see that there is almost no possibility to reject null at stage 1 ($\alpha_1 = 0.0001$) when α_0 is 0.5 for proposed BM method while BK method using Fisher's combination test still has α_1 equal to 0.0233. Actually BK method has smaller change in α_1(from 0.0233 to 0.0299) when α_1 changes from 0.3 to 0.5 than those of new method (Table 10.1), which changes from 0.00001 to 0.0335. Type I error spent at Stage 1, α_1, for both new BM method and BK Fisher's combination test are found in the same magnitude when α_0 is small and ranging from 0.30 to 0.40; and the discrepancies become larger as α_0 become large. For example α_0=0.45 and 0.5.

Table 38(Tab. 10.1): Critical values

Table 10.1: Critical values for new BM combination test as compared with BK method using Fisher's combination rule. Stage 1 critical value z_{α_1} equals to $\Phi_0^{-1}(1 - \alpha_1)$.

	New BM		BK	
α_0	α_1	Z_{α_1}	α_1	Z_{α_1}
0.30	0.0335	1.8319	0.0299	1.8817
0.35	0.0332	1.8357	0.0277	1.9163
0.40	0.0286	1.9013	0.0263	1.9380
0.45	0.0166	2.1289	0.0248	1.9642
0.50	0.00001	4.2649	0.0233	1.9896

Section 10.3: Theoretic Power, Expected Sample Size and Sample Size Re-estimation

The power of new combination test based on independent p-values from respective stages for a pre-specified alternative $\mathrm{H_A}: \frac{\mu_1-\mu_2}{\sigma} = \delta$ is:

$$\int_0^{\alpha_1} f_\delta(p_1)dp_1 +$$
$$\int_{\alpha_1}^{\alpha_0}\int_0^{A(p_1)} f_\delta(p_1,p_2)\,dp_2dp_1 \quad (10.2)$$
$$= 1-\int_{\alpha_0}^{1} f_\delta(p_1)dp_1 - \int_{\alpha_1}^{\alpha_0} f_\delta(p_1)dp_1 + \int_{\alpha_1}^{\alpha_0}\int_0^{A(p_1)} f_\delta(p_1)f_\delta(\,p_2)\,dp_2dp_1 \quad (10.3)$$
$$= 1-\int_{\alpha_0}^{1} f_\delta(p_1)dp_1 - \int_{\alpha_1}^{\alpha_0} f_\delta(p_1)dp_1 + \int_{\alpha_1}^{\alpha_0} f_\delta(p_1)\left[1-\int_{A(p_1)}^{1} f_\delta(\,p_2)\,dp_2\right]dp_1 \quad (10.4)$$
$$= 1-\int_{\alpha_0}^{1} f_\delta(p_1)dp_1 - \int_{\alpha_1}^{\alpha_0}\int_{A(p_1)}^{1} f_\delta(p_1)f_\delta(\,p_2)\,dp_2dp_1 \quad (10.5)$$

The first and second term in Equation 10.2, respectively, is the rejection probability at Stage 1 and Stage 2. Because of independence, density $f_\delta(p_1)$ can be pulled out from the inner integration in Equation 10.4. After above simplifications, the power calculation for two-stage design goes to derive individual probability densities of p_1 and p_2.

Because inverting p-value results in a standard normal, the densities of p_1 and p_2 can be derived by variable substitution. Let ϕ_δ and ϕ_0 respectively be normal density with mean δ and 0 and variance of 1. Φ_0^{-1} denotes the inverse of standard normal cumulative distribution function (CDF).

$$f_\delta(p_i) = \phi_\delta(\Phi_0^{-1}(1-p_i))d(\Phi_0^{-1}(1-p_i)) = \phi_\delta(\Phi_0^{-1}(1-p_i))|\frac{d\left(\Phi_0^{-1}(1-p_i)\right)}{dp_i}|dp_i$$
$$=\phi_\delta(\Phi_0^{-1}(1-p_i))\frac{1}{\phi_0(\Phi_0^{-1}(1-p_i))}dp_i = \frac{\phi_\delta\left(\Phi_0^{-1}(1-p_i)\right)}{\phi_0(\Phi_0^{-1}(1-p_i))}dp_i$$

When one has $N(\mu_1,\sigma^2)$ and $N(\mu_2,\sigma^2)$ for independent and identically distributed subjects within each treatment group, assuming equal size in both stages, we again accumulate n_1 and n_2 pairs of subjects at stage 1 and stage 2, respectively.

The expected sample size for this combination procedure can be easily obtained from the density function of p_1. The total expected size equals the n_1+ n_2*(Probability of continuing into Stage 2). When null hypothesis is true and $f_{\delta=0}(p_i) = 1$, the expected sample size under null hypothesis is:

$$E_{H_0}(N) = n_1 + \ n_2 * \int_{\alpha_1}^{\alpha_0} f_{\delta=0}(p_i) dp_1 = n_1 + \ n_2 * (\alpha_0 - \alpha_1) \qquad (10.6)$$

The expected sample size under alternative needs numerical integration.

$$E_{H_A}(N) = n_1 + \ n_2 * \int_{\alpha_1}^{\alpha_0} f_{\delta}(p_1) dp_1 = n_1 + \ n_2 * \int_{\alpha_1}^{\alpha_0} \frac{\phi_\delta(\Phi_0^{-1}(1-p_1))}{\phi_0(\Phi_0^{-1}(1-p_1))} dp_1 \qquad (10.7)$$

when p_1 is derived from t-test statistic.

With ratio in sample size (Stage 1 vs. total sample size) being r, then $n_1 = nr$ and $n_2 = n(1-r)$ and r=0.3, 0.5 or 0.7. With mean difference being 0.3, standard derivation being 1, one-sided type I error being 0.05, total sample size of 105 (or 137 or 190) for fixed-sample design to ensure power of 0.7 (or 0.8 or 0.9) (Table 10.2). Due to early rejection for efficacy and early stopping for futility which can possibly save sample size, one found that the expected sample sizes under all alternatives were smaller than that of the fixed sample design and were substantially reduced under null hypothesis. The theoretic power values under alternative hypotheses were as higher as or higher than respective power for fixed sample design. We also note that the overall power increases as r increases, which further suggests that the early stopping for efficacy or futility at Stage 1 makes this two-stage procedure more powerful as compared with fixed sample design because larger r allocates more subjects into Stage 1. Power also increases as α_0 increases, with which more trials stops early for futility when no sign of effect is shown at interim.

Table 39(Tab. 10.2): theoretic values of overall power and expected sample size for proposed two-stage procedure

Table 10.2: theoretic values of overall power and expected sample size for proposed two-stage procedure

		Power			EH0(N)			EHA(N)		
$(\mu_1 - \mu_2)$; σ; α n_{fixed} $1 - \beta$	$\alpha_0 =$ $\alpha_1 =$	0.3 0.0335	0.4 0.00286	0.5 0.00001	0.3 0.0335	0.4 0.00286	0.5 0.00001	0.3 0.0335	0.4 0.00286	0.5 0.00001
0.3; 1; 0.05 105 0.7	r=0.3	0.749	0.827	0.885	51.46	59.11	68.50	68.23	75.59	96.55
	r=0.5	0.845	0.901	0.939	66.86	72.31	79.00	77.35	81.65	101.70
	r=0.7	0.903	0.941	0.966	82.26	85.51	89.50	86.84	88.92	103.78
0.3; 1; 0.05 137 0.8	r=0.3	0.797	0.865	0.913	66.59	76.65	89.00	87.98	96.89	128.52
	r=0.5	0.892	0.934	0.961	87.12	94.25	103.00	98.12	102.92	134.44
	r=0.7	0.940	0.966	0.981	106.92	111.23	116.50	110.33	112.54	135.77
0.3; 1; 0.05 190 0.9	r=0.3	0.8588	0.911	0.945	92.45	106.39	123.50	118.08	128.65	182.38
	r=0.5	0.938	0.965	0.981	120.32	130.28	142.50	128.48	133.66	187.13
	r=0.7	0.973	0.986	0.993	148.19	154.17	161.50	147.09	149.21	187.96

Conditional power is defined as the probability of rejection at Stage 2, provided that the estimated treatment effect from stage 1 is carried over to Stage 2. For the case of testing mean difference for two independent normal data with known variance, null will be rejected if $p_2 \leq A(p_1)$, which is the same as $T_2 \geq z_{1-A(p_1)}$. With X and Y to indicate endpoints in treatment 1 and 2, respectively,

$$T_2 = \frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{2\sigma^2/\, n_2^*}} = \frac{\sum_{i=1}^{n_2^*}(X_i-Y_i)}{\sqrt{2\sigma^2/\, n_2^*}} = \frac{\bar{X}-\bar{Y}}{\sqrt{2\sigma^2/\, n_2^*}}$$

With assuming treatment effect observed at interim is carried forward to the final analysis, $\hat{\delta}=\frac{t_1}{\sqrt{n_1/2}}$ because of $E(T_1)= \sqrt{n_1/2}\hat{\delta}$. Therefore, the power at stage two is:

$$P_{H_A}\left(T_2 \geq z_{1-A(p_1)}\right) = P_{H_A}\left(T_2 - \sqrt{\frac{n_2^*}{2}}\hat{\delta} \geq z_{1-A(p_1)} - \sqrt{\frac{n_2^*}{2}}\hat{\delta}\right) = P_{H_A}\left(T_2 - \sqrt{\frac{n_2^*}{2}}\frac{t_1}{\sqrt{\frac{n_1}{2}}} \geq\right.$$

$$\left. z_{1-A(p_1)} - \sqrt{\frac{n_2^*}{2}}\frac{t_1}{\sqrt{\frac{n_1}{2}}}\right)=1\text{-}\Phi(z_{1-A(p_1)} - \sqrt{\frac{n_2^*}{2}}\frac{t_1}{\sqrt{\frac{n_1}{2}}})$$

Equating $1\text{-}\Phi(z_{1-A(p_1)} - \sqrt{\frac{n_2^*}{2}}\frac{t_1}{\sqrt{\frac{n_1}{2}}})$ with required power for stage 2 test of $1\text{-}\beta_2$, we can solve n_2^* for Stage 2 sample size. That is $n_2^* = n_1\frac{(z_{1-A(p_1)} + z_{1-\beta_2})^2}{t_1^2}$ (10.8)

Section 10.4: Simulations for Operating Characteristics

In Table 10.3, simulations with 100000 iterations for each scenario were used to assess type I error for proposed BM combination test. And it was shown in Table 10.3 that all simulated errors suggested that type I error was well-controlled. In Table 10.4, simulations were done to check conditional power after sample size adaptation, overall power for this BM method, average sample size at Stage 2 and average sample size for this adaptive two-stage procedure. In order to control Stage 2 sample size, constrains on both maximum and minimum were put on Stage 2 sample size, which ensured it can't be greater than 4*n_{fixed}-n_1 and can't be smaller than n_{fixed}-n_1. It is that real implemented stage sample size $\mathrm{n}_2^{\#} = \max(\min(\mathrm{n}_2^{*}, 4 * n_{fixed} - n_1), n_{fixed}$-$n_1$), where n_2^{*} is defined in Equation 8 using conditional power.

Simulations for related scenarios for BK method using Fisher's combination rule were also carried out for purpose of comparison (Table 10.5). Substantially simulation results have shown that the proposed method can be implemented in trials but with less efficiency as compared with well-known BK method using Fisher's combination rule. The rationales behind this are still unknown to us and are beyond the scope of this paper.

Table 40(Tab. 10.3): simulated Type I error for new BM combination test

Table 10.3: simulated Type I error for new BM combination test.

$(\mu_1 - \mu_2);\ \sigma;\ \alpha$	n	r	Simulated Type I error		
			$\alpha_0 = 0.3$ $\alpha_0 = 0.0335$	$\alpha_0 = 0.4$ $\alpha_0 = 0.0286$	$\alpha_0 = 0.5$ $\alpha_0 = 0.00001$
0; 1; 0.05	105	0.3	0.0502	0.0495	0.0502
		0.5	0.0482	0.0503	0.0493
		0.7	0.0506	0.0506	0.0489
	137	0.3	0.0496	0.0505	0.0490
		0.5	0.0501	0.0499	0.0503

Table 41(Tab. 10.4): simulated values of overall power and expected sample size

Table 10.4: simulated values of overall power and expected sample size for proposed two-stage procedure

		BM method Conditional Power (Stage 2) / Overall power(two stages) ASN (Stage 2) / ASN(two stages)		
$(\mu_1 - \mu_2)$; σ; α n_{fixed} $1-\beta$	$\alpha_0 =$ $\alpha_1 =$	0.3 0.0335	0.4 0.00286	0.5 0.00001
0.3; 1; 0.05 105 0.7	r=0.3	0.7268/0.5658 192/177	0.7004/0.6528 199/236	0.7717/0.6827 232/201
	r=0.5	0.7592/0.6720 192/165	0.7479/0.7654 193/234	0.7839/0.7365 223/171
	r=0.7	0.7953/0.7579 201/161	0.7927/0.8447 194/238	0.7874/0.7631 221/149
0.3; 1; 0.05 137 0.8	r=0.3	0.7962/0.6593 233/217	0.7699/0.7395 233/294	0.8314/0.7609 282/241
	r=0.5	0.8199/0.7629 226/196	0.8107/0.8417 221/289	0.8379/0.8070 262/194
	r=0.7	0.8505/0.8391 231/189	0.8494/0.9054 218/295	0.8353/0.8221 255/162
0.3; 1; 0.05 190 0.9	r=0.3	0.8698/0.7659 288/176	0.8354/0.8237 278/381	0.8941/0.8463 349/294
	r=0.5	0.8909/0.8654 265/236	0.8703/0.9124 249/372	0.9000/0.8838 312/220
	r=0.7	0.9071/0.9193 265/224	0.9060/0.9695 241/379	0.8807/0.8789 296/169

Table 42(Tab. 10.5): simulated values of overall power and expected sample

Table 10.5: Simulated values of overall power and expected sample size for BK method using Fisher's combination rule

		BK method Conditional Power (Stage 2) / Overall power(two stages) ASN (Stage 2) / ASN(two stages)		
$(\mu_1 - \mu_2)$; σ; α n_{fixed} $1 - \beta$	$\alpha_0 =$ $\alpha_1 =$	0.3 0.0299	0.4 0.0263	0.5 0.0233
0.3; 1; 0.05 105 0.7	r=0.3	0.9014/0.7013 168/205	0.9127/0.7771 193/225	0.9199/0.8311 210/236
	r=0.5	0.9407/0.8196 185/234	0.9421/0.8695 198/239	0.9410/0.9026 210/242
	r=0.7	0.9651/0.8881 191/248	0.9634/0.9248 201/247	0.9573/0.9437 208/245
0.3; 1; 0.05 137 0.8	r=0.3	0.9406/0.7728 200/260	0.9477/0.8382 230/282	0.9509/0.8833 248/293
	r=0.5	0.9652/0.8762 213/293	0.9648/0.9159 229/297	0.9633/0.9410 237/297
	r=0.7	0.9814/0.9317 216/312	0.9801/0.9580 226/309	0.9760/0.971 233/303
0.3; 1; 0.05 190 0.9	r=0.3	0.9707/0.8465 240/338	0.9732/0.8963 273/363	0.9755/0.9304 292/376
	r=0.5	0.9830/0.9343 241/380	0.9815/0.9581 259/383	0.9822/0.9721 268/379
	r=0.7	0.9927/0.710 242/409	0.9915/0.9830 250/399	0.9899/0.9895 256/385

Section 10.5: Discussion

Similar to BK method using Fisher's combination rule, proposed new BM method combines p-values from two disjoint samples together to form a two-stage adaptive procedure. The validity of this method inherits from distributional property of this combination function of two independent p-values, along with formulas to calibrate conditional error for Stage 2 to ensure overall type I error control. Type I error is well-controlled based on asymptotical theory and then further confirmed by simulation results. Operational characteristics in terms of power and expected sample size under null and alternative hypotheses were also shown for this new BM combination test as compared with BK method using Fisher's combination rule. Due to the invariance of p-value to be uniformly distributed from 0 to 1, this method can be applied all data type as long as p-values are from disjoint samples or independent asymptotically.

References:

P. Bauer and K. Köhne, 1994. Evaluation of Experiment with Adaptive Interim Analysis. *Biometrics*, 50: 1029-1041.

Michael A. Proschan and Sally A. Hunsberger, 1995. Designed Extension of Studies Based on Conditional Power. *Biometrics*, 51:1315-1324.

Gernot Wassmer, 1998. A Comparison of Two Methods for Adaptive Interim Analysis in Clinical Trials. *Biometrics*, 54:696-705.

Martin Posch and Peter Bauer, 1999. Adaptive Two Stage Designs and the Conditional Error Function. *Biometrical Journal*, 41:689-696

Janet Wittes and Erica Brittain. 1990. The Role of Internal Pilot Studies in Increase the Efficiency of Clinical Trials. *Statistics in Medicine*, 9:65-72.

Gould, A. L. 1992. Interim Analysis for Monitoring Clinical trials that do not materially affect the type I Error Rate. *Statistics in Medicine*, 11:55-66.

Liu, Qing and George Y. H. Chi. 2001. On Sample Size and Inference for Two-stage Adaptive Designs. *Biometrics*, 57:172-177.

Walter Lehmacher and Gernot Wassmer. 1999. Adaptive Sample Size Calculations in Group Sequential Trials. Biometrics, 55:1286-1290.

Box, G.E.P. and Muller, M.E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.*, 29:610-611