

Thomas C. Chalmers Scholarship Finalists

2026

Author: Xi Fang, PhD
Institute: Yale School of Public Health
Title: Sample Size Determination for Win Statistics in Cluster-Randomized Trials

Abstract:

Composite endpoints are increasingly used in clinical trials to capture treatment effects across multiple or hierarchically ordered outcomes. Although inference procedures based on win statistics—such as the win ratio, win odds, and net benefit—have gained traction in individually randomized trials, their methodological development for cluster-randomized trials remains limited. In particular, there is no formal framework for power and sample size determination when using win statistics with composite time-to-event outcomes. We develop a unified framework for power and sample size calculation for win statistics under cluster randomization. Analytical variance expressions are derived for a broad class of win statistics, yielding closed-form variance expressions and power procedures that avoid computationally intensive simulations. The variance expressions explicitly characterize the roles of the rank intracluster correlation coefficient, cluster size, tie probability, and outcome prioritization for study planning purposes. Importantly, our variances nest existing formulas for univariate outcomes as special cases while extending them to complex, hierarchically ordered composite endpoints. Simulation studies confirm accurate finite-sample performance, and we supply a case study to illustrate the use of our method to re-design a real-world cluster-randomized trial.

Author: Yunyi Liu (PhD Student)
Institute: UC San Diego
Title: Bayesian Response-Adaptive Randomization for Cluster Randomized Controlled Trials

Abstract:

Cluster randomized controlled trials where groups (or clusters) of individuals, rather than single individuals, are randomized are especially useful when individual-level randomization is not feasible or when interventions are naturally delivered at the group level. Balanced randomization in the cluster randomized trial setting can pose logistical challenges and strain resources if subjects are randomized to a non-optimal arm. We propose a Bayesian response-adaptive randomization design for cluster randomized controlled trials based on Thompson sampling, which dynamically allocates clusters to the most efficacious treatment arm based on the interim posterior distributions of treatment effects using Markov chain Monte Carlo sampling. Our design also incorporates early stopping rules for efficacy and futility determined by prespecified posterior probability thresholds. The performance of the proposed design is evaluated across various operating characteristics under multiple settings, including varying intra-cluster correlation coefficients, cluster sizes, and effect sizes. Our adaptive approach is also compared with a standard, parallel two-arm cluster randomized controlled clinical trial design, highlighting improvements in both ethical considerations and efficiency. From our simulation studies based on an HIV behavioral trial, we demonstrate these improvements by preferentially assigning more clusters to the more efficacious intervention while maintaining robust statistical power and controlling false positive rates.

Author: Zizhong Tian, PhD (Winner)
Institute: Eli Lilly and Company
Title: Bayesian Meta-Analysis of Comparative Drug Safety

Abstract:

Meta-analysis is a fundamental tool for synthesizing evidence from clinical trials. Beyond efficacy assessment, meta-analyses of safety data pose distinct methodological challenges. Particularly, adverse drug reactions are often rare, multi-dimensional, on an ordinal severity scale, and incompletely reported, rendering traditional meta-analysis methods and assumptions developed for assessing treatment efficacy inappropriate for comparing drug safety. To address these challenges, we introduce a Bayesian meta-analysis approach to integrating complex AE data for comprehensive safety assessment. The proposed modeling strategy jointly analyzes multivariate AEs and their ordinal severity information, incorporates sparsity-inducing priors to characterize AE-specific treatment discrepancies, and accounts for left-censoring due to incomplete reporting, collectively providing reliable and

efficient estimation of AE incidence probabilities and odds ratios in synthesizing the (comparative) drug safety evidence across a combination of two-armed randomized clinical trials and single-armed studies. Simulation studies demonstrate the model's robustness to rare AE events and estimation accuracy of AE risks with effective false discovery control in comparative safety signal detection. A motivating example in targeted cancer therapy further illustrates the model's practical utility and interpretability in drug safety evaluation. With wide application potential, the proposed general-purpose method will enable evidence-based medicine in the current data-rich era by better characterizing the safety profile of new medical interventions.

2025

Author: Lee Ding
Institute: Harvard University
Title: Power Calculation for Group Sequential Cluster Randomized Trials With Continuous or Binary Outcomes

Abstract:

Well-planned interim analyses provide researchers with early data, allowing for timely decisions about a trial's continuation, modification, or termination to enhance patient safety, reduce costs, and expedite access to effective treatments. Group sequential methods enable investigators to stop a clinical trial early when there is compelling evidence for efficacy or futility while preserving the trial's statistical integrity. Although group sequential methods are well developed for individually randomized trials, established methods for cluster randomized trials (CRTs) are limited. In CRTs, groups or clusters (such as communities, schools, or clinics) rather than individual participants are randomly assigned to treatment or control conditions, making this design especially useful in settings where individual randomization is impractical or when the intervention is delivered at a group level. Because clusters are defined based on some shared characteristics or circumstances, outcomes for individuals within the same cluster tend to be more similar than those in different clusters. Group sequential trials require an inflated maximum sample size compared to equivalent fixed-sample designs to account for the possibility of early stopping. However, limited guidance exists on designing and powering a group sequential CRT, which requires accounting for correlated outcomes and repeated interim analyses of data accumulating at both the cluster and individual levels.

To this end, we develop sample size calculation methods for group sequential CRTs with continuous or binary endpoints. Under designs that recruit by cluster or individuals within clusters, we first show that differences between the corresponding sequentially calculated test statistics are asymptotically independent. We then employ an error spending approach to determine the maximum number of clusters or cluster size of a trial. Our method encompasses early stopping for combinations of efficacy and binding or non-binding futility. In simulation studies, we find that group sequential CRTs powered using our sample size calculations achieve the specified power across a range of trial design specifications; these results hold even when both clusters and individual participants enter the trial at varying levels over time. We also provide guidance on how and when to schedule interim analyses to maximize the efficiency of a group sequential CRT. We then apply our approach to planning interim analyses in the MEDUSA study, a CRT evaluating the effect of a multifaceted antimicrobial therapy initiation program on sepsis survival.

Author: Auden Krauska
Institute: University of Wisconsin-Madison
Title: Assessing Adherence to CONSORT Reporting Guidelines Using AI

Abstract:

Background: The CONSORT statement provides recommendations in the form of a checklist for comprehensive and transparent reporting of randomized controlled trials (RCTs). However, verifying adherence to the checklist is time-consuming. Large Language Models (LLMs) could automate this process.

Methods: We analyzed 10 RCTs published in the BMJ, comparing page numbers recorded by authors in completed CONSORT checklists with those identified by the Rohe Nordberg CONSORT Report, a multi-stage LLM system. Agreement was assessed when page numbers overlapped and analyzed separately for fully, partially, and incompletely reported items. A random sample of disagreements underwent human review.

Results: Agreement varied by reporting category: 72.7% for items marked not applicable, 56.1% for fully reported items, 45.6% for partially reported items, and 44.7% for incompletely reported items. Paper-level agreement ranged from 2.7% to 95%. Human review of disagreements found that in 29% of cases, both sources were valid but different; in 29%, the Rohe Nordberg CONSORT Report identifications pertained more to the given item, while the author citations pertained somewhat; and in 43%, author citations were not consistent with the checklist item while the Rohe Nordberg CONSORT Report's paragraph citations were.

Conclusions: LLMs show promise in automating adherence checking of CONSORT, with high agreement for basic trial information and ability to identify non-applicable items. Lower agreement for complex methodological items and wide variation across papers suggest areas for author education. The LLM were more precise in locating the relevant page numbers than authors, indicating potential value for improving reporting quality.

Author: Mollie Payne, MSc (Winner)
Institute: King's College London
Title: Dose Outcome Using Stratified Estimation With Random Forest Method (DOSE-RF): A Novel Approach to Dose-Response Modeling in Complex Interventions

Abstract:

Background: In randomised controlled psychotherapeutic trials, we often ask: What is the effect of being offered therapy? However, without perfect compliance, this differs from assessing the effect of receiving therapy. Traditional approaches that only analyse dose-response within the treatment arm can lead to biased, non-causal estimates. Newer methods address these biases but rely on prior assumptions about the dose-response relationship, typically assuming linearity. We propose the DOSE-RF method, a novel, causally valid approach that overcomes these limitations by estimating a dose-response function without predefined assumptions.

Methods: The DOSE-RF method combines machine learning and principal stratification in a two-stage procedure to estimate the dose-response effect. In stage one, random forest classification predicts counterfactual dose for the control arm. In stage two, regression models calculate the causal effect within each dose level. We tested this method across 36 simulation scenarios, varying dose predictors, confounding, dose distribution, and the dose-response function. The method was applied to two illustrative examples: the SoCRATES trial, a trial aimed at reducing paranoia in patients with schizophrenia, and the COMPASS trial, which focused on alleviating psychological distress in patients with long-term health conditions.

Results: In simulations, the DOSE-RF method reliably detected the true dose-response function across diverse scenarios, achieving accurate results without requiring prior assumptions. For scenarios with normally distributed doses, the method showed no significant bias. In cases where the dose followed a beta distribution, some bias emerged due to strong unmeasured confounding, particularly in small sample sizes. Application to the SoCRATES trial confirmed these patterns, though estimation at lower dose levels was challenging due to limited data. Despite this, DOSE-RF successfully estimated stratified treatment effects across most dose levels and identified the number of sessions needed to significantly decrease paranoia in patients. In the COMPASS trial, DOSE-RF revealed heterogeneous treatment effects, which would enable researchers to pinpoint both the minimum effective dose and the optimal next-best dose for patients who may require more than the minimum. These findings highlight DOSE-RF's potential to support more personalised and effective treatment recommendations in clinical practice.

Conclusions: The DOSE-RF method offers a robust and flexible approach for analysing dose-response relationships in clinical trials, addressing biases from non-compliance and restrictive assumptions regarding the true dose-response function. Application to real-world examples highlight the possibilities of this method and allow us a deeper insight into the effect of therapy. It is particularly suited for trials with larger sample sizes and a moderate number of dose levels. Trialists should consider factors influencing compliance and include these in their data collection to optimise the method's performance.

2024

Author: Jessica Wild
Institute: University of Colorado Anschutz Medical Campus
Title: Upstrapping to Determine Fertility: Predicting Future Outcomes Nonparametrically From Past Data

Abstract:

Background: Clinical trials often involve some form of interim monitoring to determine futility before planned trial completion. While many options for interim monitoring exist (e.g., alpha-spending, conditional power, etc.), nonparametric based interim monitoring methods are also needed to account for more complex trial designs and analyses. The upstrap is one recently proposed nonparametric method that may be applied for interim monitoring. **Methods:** Upstrapping is motivated by the case resampling bootstrap and involves repeatedly sampling with replacement from the interim data to simulate thousands of fully enrolled trials. The p-value is calculated for each upstrapped trial and the proportion of upstrapped trials for which the p-value criteria are met is compared with a pre-specified decision threshold. To evaluate the potential utility for upstrapping as a form of interim futility monitoring, we conducted a simulation study considering different sample sizes with several different proposed calibration strategies for the upstrap. We first compared trial rejection rates across a selection of threshold combinations to validate the upstrapping method. Then we applied upstrapping methods to simulated clinical trial data, directly comparing their performance with more traditional alpha-spending and conditional power interim monitoring methods for futility.

Results: The method validation demonstrated that upstrapping is much more likely to find evidence of futility in the null scenario than the alternative across a variety of simulations settings. Our three proposed approaches for calibration of the upstrap had different strengths depending on the stopping rules used. Compared to O'Brien-Fleming group sequential methods, upstrapped approaches had type I error rates that differed by at most 1.7% and power that was between 15.7% lower and 0.2% higher, while expected sample size was 2-22% lower in the null scenario

Conclusions: In this proof-of-concept simulation study, we evaluated the potential for upstrapping as a resampling-based method for futility monitoring in clinical trials. The trade-offs of the expected sample size, power, and type I error rate control indicate that different upstrap calibrations achieve more or less aggressive stopping for futility and that performance similarities can be identified relative to the considered alpha-spending and conditional power methods for futility monitoring.

Author: Peter Zhang (Winner)
Institute: Medical College of Wisconsin
Title: Covariate-Adjusted Group Sequential Comparisons of Survival Probabilities

Abstract:

In clinical trials, survival outcomes are frequently studied and interim analyses are often desirable. Common analysis methods such as the log rank test and Cox regression model rely on a proportional hazards (PH) assumption and are subject to type I error rate inflation and loss of power when PH are violated. Such violations may be expected a priori, particularly when the mechanisms of treatments differ such as immunotherapy vs. chemotherapy for treating cancer. We develop covariate-adjusted group sequential (GS) tests for comparing survival probabilities that permit non-PH between treatments and interim analyses for efficacy and/or futility, offering easily interpretable and clinically meaningful summary measures of the treatment effect. The test statistic sequences converge to the canonical joint distribution, facilitating selection of critical values and sample size for a GS trial to meet type I error rate and power requirements. Simulations demonstrate that the type I error rate and power of the proposed tests meet targeted levels and are robust to both the PH assumption and covariate influence. The proposed tests are illustrated using data from the BMT CTN 1101 clinical trial.

Author: Kehao Zhu
Institute: University of Washington
Title: Designing Cancer Screening Trials for Reduction in Late-Stage Cancer Incidence

Abstract:

Before implementing a biomarker test for cancer screening into routine clinical care, the test must demonstrate clinical utility. Unlike therapeutic trials for patients diagnosed with cancer, designing a randomized controlled trial (RCT) to demonstrate the clinical utility of an early detection biomarker test with mortality and related endpoints poses unique challenges. The hurdles stem from the prolonged natural progression of the disease and the lack of information regarding the time-varying screening effect on the targeted asymptomatic population. To facilitate the

study design of cancer screening trials, we propose using a generic multistate disease history model and derive model-based effect sizes. The model links key performance metrics of the test to primary endpoints like the incidence of late-stage cancer. Based on the chronological time scale aligned with the RCT, our method allows the assessment of study powers based on key design features of the new program, including the test sensitivity, the number and frequency of repeated tests, and the length of follow-up. We have made a tool from the proposed method available as an easy-to-use and transparent web application. The tool will enable trialists and biomarker researchers to perform realistic and quick evaluations when strategizing screening trials for specific diseases. We use numerical examples based on the National Lung Screening Trial (NLST) to demonstrate the novel method.

2023

Author: Xuetao Lu
Institute: The University of Texas MD Anderson Cancer Center
Title: Distribution-free Overlapping Indices for Efficient and Robust Information Borrowing in Bayesian Hierarchical Modeling

Abstract:

Bayesian hierarchical model (BHM) has been widely used in synthesizing information across subgroups in clinical trials. Identifying heterogeneity in the data and determining the proper strength of borrow have long been crucial goals pursued by researchers. However, because of their internal connections, we must consider them together. This joint consideration presents two fundamental challenges: (1) How can we balance the trade-off between homogeneity within cluster and adaptability for information borrowing? (2) How can we dynamically borrow information in different clusters? To tackle these challenges, we first propose two novel distribution-free overlapping indices: the overlapping clustering index (OCI) for identifying the optimal clustering result and the overlapping borrowing index (OBI) for assigning proper borrowing strength to clusters. A weighted K-Means clustering method, equivalent to maximizing OCI, is developed to perform optimal clustering. Subsequently, we construct a new method BHMOI (Bayesian hierarchical model with overlapping indices) by embedding OCI and OBI into the BHM framework. BHMOI can achieve efficient and robust information borrowing with desirable properties. Examples and simulation studies are provided to demonstrate the effectiveness of BHMOI in heterogeneity identification and dynamic information borrowing.

Author: Sidi Wang (2023 Winner)
Institute: University of Michigan
Title: Dynamical Enrichment of Bayesian Small Sample, Sequential, Multiple Assignment Randomized Trial (snSMART) Design Using Natural History Data: A Case Study From Duchenne Muscular Dystrophy

Abstract:

In Duchenne muscular dystrophy (DMD) and other rare diseases, recruiting patients into clinical trials is challenging. Additionally, assigning patients to long-term, multi-year placebo arms raises ethical and trial retention concerns. This poses a significant challenge to the traditional sequential drug development paradigm. In this article, we propose a small sample, sequential, multiple assignment, randomized trial (snSMART) design that combines dose selection and confirmatory assessment into a single trial. This multi-stage design evaluates the effect of multiple doses of a promising drug versus placebo. In stage 1, participants are randomized in greater proportion to receive low dose or high dose over placebo. In stage 2, participants are re-randomized across treatments depending on their stage 1 dose and response. Our proposed approach increases the efficiency of treatment effect estimates by i) enriching the placebo arm with external control data, and ii) using data from all stages. Data from external control and different stages are combined using a robust Meta-analytic Combined (MAC) approach to consider the various sources of heterogeneity and potential selection bias. We reanalyze data from a DMD trial using the proposed method and external control data from the Duchenne Natural History Study (DNHS). Our method's estimators show improved efficiency compared to the original trial. Also, the robust MAC-snSMART provides more accurate estimators than the traditional analytic method when its assumptions (practical in most snSMART regimes) are not violated. Overall, the proposed methodology provides a promising candidate for efficient drug development in DMD and other rare diseases.

Author: Chao Yang
Institute: The University of Texas MD Anderson Cancer Center
Title: An Extended Bayesian Semi-Mechanistic Dose-Finding Design for Phase I Oncology Trials Using

Abstract:

The primary aims of a phase I oncology trial are to evaluate the safety profile of an investigational drug and identify the maximum tolerated dose (MTD) or maximum tolerated dose-regimen (MTD-regimen). Standard dose-finding designs do not systemically take pharmaco-kinetic/dynamic (PK/PD) information into account when modeling the dose-toxicity relationship. We propose a model-based, semi-mechanistic dose-finding (SDF) design that incorporates relevant PK/PD information to model the dose-toxicity relationship. This design extends a recently proposed SDF model framework, which uses sequentially a population PK model for the concentration-time profile, a PD model that maps drug concentration to a PD effect, and a link function that associates the cumulative PD effect with the DLT probability, to incorporate measurements for a PD biomarker relevant to the primary dose-limiting toxicity (DLT). We propose a joint Bayesian modeling of the PK, PD, and DLT outcomes. As such, the effect of a dose/regimen on its associated DLT probability is modeled through drug exposure and cumulative PD effect, which depends on PK and PD parameters. We also introduce a parameter in the link function to allow for more flexibility in its modeling.

Our extensive simulation studies show that on average our proposed design outperforms several common phase I trial designs, including the modified toxicity probability interval (mTPI) and Bayesian optimal interval (BOIN) designs, the continual reassessment (CRM) method, as well as an SDF design assuming a latent PD biomarker, in terms of the probability of correct selection of MTD and average number of patients allocated at MTD under a variety of dose-toxicity scenarios. The proposed design also yields better estimated dose-toxicity curves than CRM designs in scenarios where an MTD exists. In a sensitivity analysis, we find the performance of the proposed design is robust to prior specification for the parameter in link function. When the prior mean departs moderately from the truth or when the prior variance is large, the proposed design still yields adequate (better or comparable) performance compared to the competing designs.

2022

Author: Peter Godolphin (2022 Winner)

Institute: University College of London

Title: Estimating Interactions and Subgroup-Specific Treatment Effects in Meta-Analysis Without Aggregation Bias: A Within-Trial Framework

Abstract:

Estimation of within-trial interactions in meta-analysis is crucial for reliable assessment of how treatment effects vary across participant subgroups. However, currently-available accessible methods have various limitations: they mostly focus on covariates with only two subgroups, and may exclude relevant data if only a single subgroup is reported. Moreover, patients, clinicians and policy-makers need reliable estimates of treatment effects within specific covariate subgroups, on relative and absolute scales, in order to target treatments appropriately – which estimation of an interaction effect does not in itself provide. Therefore, in this presentation we further develop the “within-trial” framework by providing practical methods to (1) estimate a set of within-trial interactions for any categorical covariate with two or more groups; and (2) estimate a set of “floating” subgroup-specific treatment effects that are compatible with the within-trial interactions, whilst allowing inclusion of all subgroup data. We show how these floating estimates avoid the risk of introducing aggregation bias. We implement this methodology using examples taken from previously published meta-analyses and demonstrate a straightforward implementation in Stata based upon existing code for multivariate meta-analysis. These methods can be applied using observed effect sizes and standard errors within subgroups within trials. Thus, the within-trial framework can be utilised with aggregate (or “published”) source data, as well as with individual patient data, providing wide practical application. We also suggest novel improvements to the traditional forest plot for best presenting within-trial interactions and floating subgroups.

Author: Jiyang Wen

Institute: Johns Hopkins Bloomberg School of Public Health

Title: Simultaneous Hypothesis Testing for Multiple Competing Risks in Comparative Clinical Trials

Abstract:

Competing risks data are commonly encountered in randomized clinical trials or observational studies. Ignoring competing risks in survival analysis leads to biased risk estimates and improper conclusions. Often, one of the competing events is of primary interest and the rest competing events are handled as nuisance. These approaches

can be inadequate when multiple competing events have important clinical interpretations and thus of equal interest. For example, in COVID-19 in-patient treatment trials, the outcomes of COVID-19 related hospitalization are either death or discharge from hospital, which have completely different clinical implications and are of equal interest, especially during the pandemic. In this paper we develop nonparametric estimation and simultaneous inferential methods for multiple cumulative inference functions (CIFs) and corresponding restricted mean times. Based on Monte Carlo simulations and a data analysis of COVID-19 in-patient treatment clinical trial, we demonstrate that the proposed method provides global insights of the treatment effects across multiple endpoints.

Author: Jack Wolf

Institute: University of Minnesota School of Public Health

Title: A Permutation Procedure to Detect Heterogeneous Treatments Effects in Randomized Clinical Trials While Controlling the Type-I Error Rate

Abstract:

Background/Aims: Secondary analyses of randomized clinical trials often seek to identify subgroups with differential treatment effects. These discoveries can help guide individual treatment decisions based on patient characteristics and identify populations for which additional treatments are needed. Traditional analyses require researchers to prespecify potential subgroups to reduce the risk of reporting spurious results. There is a need for methods that can detect such subgroups without a priori specification while allowing researchers to control the probability of falsely detecting heterogeneous subgroups when treatment effects are uniform across the study population.

Methods: We propose a permutation procedure for tuning parameter selection that allows for Type-I error control when testing for heterogeneous treatment effects framed within the Virtual Twins procedure for subgroup identification. We verify that the Type-I error rate can be controlled at the nominal rate and investigate the power for detecting heterogeneous effects when present through extensive simulation studies. We apply our method to a secondary analysis of data from a randomized trial of very low nicotine content cigarettes.

Results: In the absence of Type-I error control, the observed Type-I error rate for virtual twins was between 99 and 100%. In contrast, models tuned via the proposed permutation were able to control the Type-I error rate and detect heterogeneous effects when present. An application of our approach to a recently completed trial of very low nicotine content cigarettes identified several variables with potentially heterogeneous treatment effects.

Conclusions: The proposed permutation procedure allows researchers to engage in secondary analyses of clinical trials for treatment effect heterogeneity while maintaining the Type-I error rate without pre-specifying subgroups.

2021

Author: Subodh R Selukar (2021 Winner)

Institute: Department of Biostatistics, University of Washington

Title: Stratified randomization for platform trials with differing experimental arm

Abstract:

Platform trials facilitate efficient use of resources by comparing multiple experimental agents to a common standard of care arm. They can accommodate a changing scientific paradigm within a single trial protocol by adding or dropping experimental arms - critical features for trials in rapidly developing disease areas such as COVID-19 or cancer therapeutics. However, in these trials, efficacy and safety issues may render certain participant subgroups ineligible to some experimental arms, and methods for stratified randomization do not readily apply to this setting of differing experimental arm eligibility. We motivate this setting with the LEAP trial, a platform trial for acute myeloid leukemia in older adults. When experimental arms differ in eligibility, existing methods for stratified randomization require changes in trial-wide eligibility, which affects trial accrual and generalizability. This work describes how to extend conventional randomization methods to account for varying experimental arm eligibility. We suggest modifying block randomization by including experimental arm eligibility as a stratifying variable, and we suggest modifying the imbalance score calculation in dynamic balancing by performing pairwise comparisons between each eligible experimental arm and standard of care arm participants eligible to that experimental arm. We also briefly discuss the impact of differing eligibility on the efficiency of platform trials as measured by the size of the common standard of care arm.

Author: Xiaoyu Tang
Institute: Department of Biostatistics, Boston University
Title: Bayesian multivariate network meta-analysis for the difference in restricted mean survival times

Abstract:

Network meta-analysis (NMA) is essential for clinical decision-making. NMA enables inference for all pair-wise comparisons between interventions available for the same indication, by using both direct evidence and indirect evidence. In randomized trials with time-to-event outcome data, such as lung cancer data, conventional NMA methods rely on the hazard ratio and the proportional hazards assumption, and ignore the varying follow-up durations across trials. We introduce a novel multivariate NMA model for the difference in restricted mean survival times (RMST). Our model synthesizes all the available evidence from multiple time points simultaneously and borrows information across time points through within-study covariance and between-study covariance for the differences in RMST. We derived the within-study covariance and estimated the model under the Bayesian framework. We evaluated our model by conducting a simulation study. Our multiple-timepoint model yields lower mean squared error over the conventional single-timepoint model at all time points, especially when the availability of evidence decreases. We illustrated the model on a network of randomized trials of second-line treatments of advanced non-small-cell lung cancer. Our multiple-timepoint model yielded increased precision and detected evidence of benefit at earlier timepoints as compared to the single-timepoint model. Our model has the advantage of providing clinically interpretable measures of treatment effects.

Author: Siyun Yang
Institute: Biostatistics and Bioinformatics, Duke University
Title: Covariate adjustment in subgroup analyses of randomized clinical trials: A propensity score approach

Abstract:

Background: Subgroup analyses are frequently conducted in randomized clinical trials to assess evidence of heterogeneous treatment effect across patient subpopulations. Although randomization balances covariates within subgroups in expectation, chance imbalance may be amplified in small subgroups and harm the precision of subgroup analyses. Two main approaches for covariate adjustment include analysis of covariance (ANCOVA) and propensity score weighting in RCTs. In this article, we develop propensity score weighting methodology to improve the precision and power of subgroup analyses by eliminating chance imbalances.

Methods: We extend the propensity score weighting methodology to subgroup analyses by fitting a logistic regression propensity model with covariate-subgroup interactions. We show that overlap weighting exactly balances the covariates with interaction terms in each subgroup. Extensive simulations are performed to compare the operating characteristics of unadjusted estimator, different propensity score weighting estimators and the ANCOVA estimator. We apply these methods to the HF-ACTION trial to evaluate the effect of exercise training on 6-minute walk test in several pre-specified subgroups.

Results: Efficiency of the adjusted estimators is higher than that of the unadjusted estimator. The propensity score weighting estimator is as efficient as ANCOVA, and may be more efficient when subgroup sample size is small ($N < 125$), or when outcome model is mis-specified. The weighting estimators with full-interaction propensity model consistently outperform traditional main-effect propensity model.

Conclusion: Propensity score weighting serves as a transparent alternative to adjust important covariates in subgroup analyses of RCTs. It is important to include the full set of covariate-subgroup interactions in the propensity score model.

2020

Author: Thevaa Chandereng (2020 Winner)
Institute: Department of Biostatistics and Medical Informatics, University of Wisconsin
Title: Robust blocked response-adaptive randomization designs

Abstract:

1 Introduction

1.1 Response-adaptive randomization

Randomization remains a pivotal methodology for advancement in medical knowledge properly done. Traditionally, a fixed randomization scheme (usually 1:1 or 2:1) is used due to simplicity in design and execution of the trial. However, response-adaptive randomization (RAR) designs utilize accrual information to adaptively tilt the randomization ratio to the better performing treatment group. However, in traditional RAR confounding of treatment with time induces a potentially severe bias [1,13,2,8]. The purpose of this article is to expand on the characteristics of blocked RAR, proposed by Karrison et al. as a way to eliminate this bias [8]. Although, in this paper we focus on trials with two parallel intervention groups, our method are easily extendable to three or more arms.

On the other hand, opponents of RAR have argued that adaptive randomization challenges the whole notion of equipoise [1]. Hey and Kimmelman also argued that most new treatments offer small improvement over standard treatments, thus they offer limited benefit and require a larger sample size [6]. Hey and Kimmelman also suggested that equal randomization helps reduce the trial size and length, thus it benefits future patients rather than current patients enrolled in the trial [6]. Korn and Friedlin measure the difference in non-responders under equal and adaptive randomization and found that adaptive randomization required a larger trial to achieve the same power and type-I error [9]. Also, outcomes in RAR trials must be short to be able to obtain the outcome of the trial for future randomization [8].

1.2 Time-trend issues

As stated above, a major criticism of RAR is the time-trend issue. This is a main factor for why RAR is infrequently used. The type-I error rate is usually not controlled at the nominal level under traditional Bayesian or frequentist RAR designs [13]. Besides affecting type-I error, studies have shown that there is a large bias in the estimation of treatment difference under traditional RAR designs [13].

In long duration trials, time-trends are especially likely to occur. Patients' characteristics might be completely different throughout the trial or even at the beginning and end of the trial (which is also known as "patient drift") [8]. However, standard RAR analyses assume that the sequence of patients who arrive for entry into the trial represents samples drawn at random from two homogenous populations, with no drift in the probabilities of success [1,8]. This assumption is usually violated. For example, there were more smokers enrolled in the latter part of the trial than the beginning of the trial in the Lung Cancer Elimination (BATTLE) [10]. Kalish and Begg (1987) noted that in a sampling of large randomized Eastern Cooperative Oncology Group trials moderate time-trends in overall outcomes are common [7].

Time-trend can not only greatly bias the estimated in treatment effect but it can also wrongly reject a true null hypothesis. We propose a block (group-sequential) design where the randomization ratio is altered in a block level instead of a patient by patient basis using both frequentist and Bayesian approaches. The randomization ratio is kept constant in each block. The block design is similar to the stratified group design introduced by Karrison et al. [8]. We further study the robustness in different block sizes using both frequentist and Bayesian approach. We also compare these results with traditional RAR design and with fixed (1:1) randomization.

Author: Huaqing Jin

Institute: Department of Statistics and Actuarial Science, The University of Hong Kong

Title: Bayesian enhancement two-stage design with error control for phase II clinical trials

Abstract:

The phase II clinical trial is an essential and fundamental step to assess the preliminary information on drug efficacy. The goals of such trials are to screen out non-promising drugs and carry promising drugs into phase III clinical trials that are typically large-scale, expensive and time-consuming. Currently, the most popular single-arm phase II clinical trial design is proposed by Simon (1989) which is based on a hypothesis testing framework. Following Simon's design, there are abundant variations and extensions. (Ensign et al., 1994, Shuster, 2002, Lin and Shih, 2004, Chen and Shan, 2008, Shan et al., 2016).

However, these designs are criticized by their failure in screening out the non-effective drugs for subsequent large-scale phase III trials (Van Norman, 2019). Gan et al. (2012) investigated 235 phase III randomized cancer trials published in 10 medical journals and found that only 38% of them achieved significant results. The main reason for

such a high failure rate is the existence of the indifference region between the null and alternative hypotheses in Simon's two-stage design. Because of the indifference region, rejecting the null hypothesis does not mean that the drug achieves the target clinical response rate.

Shi and Yin (2018) proposed the Bayesian Enhancement Two-stage (BET) design to address such issue. The BET design is also built upon the hypotheses, $H_0 : p \leq p_0$ vs $H_1 : p \geq p_1$; where p is the response rate of the drug, p_0 is the clinical uninteresting response rate and p_1 represents the desirable target response rate. The BET design is characterized by four parameters (r_1, n_1, r, n) via the posterior probabilities of H_0 and H_1 and the highest posterior density (HPD) intervals. Let y_1 and y_2 denote the numbers of responses observed in the first and second stages, respectively. In the first stage, the sample size is n_1 and if $y_1 \geq r_1$, the trial would proceed to the second stage, otherwise the trial is terminated early for futility. In the second stage, $n_2 = n - n_1$ new subjects are enrolled. If at the end of the trial the total number of responses $y = y_1 + y_2$ reaches r , the drug is considered as promising; otherwise, the drug is announced as non-promising.

The BET design renders a good control of the posterior probability of H_0 when carrying the trial to the second stage and that of H_1 when declaring the drug as promising. However, from an intuitive and practical perspective, the length of HPD interval lacks transparency and interpretability, and thus the related design parameters (ℓ_1, ℓ_2) do not have a clear range to choose from. To circumvent this problem, we adapt the concepts, posterior false positive and false negative error rates in Lee and Zelen (2000), which are the counterparts of type I and type II error rates in the Bayesian framework. Based on these concepts, we replace the constraints on HPD interval lengths with posterior error probabilities when rejecting the drug at stage 1 and stage 2. Unlike the BET design which mainly focuses on reducing posterior error rates under the minimal required response number and uses lengths of HPD intervals to control the variance, we propose the BET design with error control by explicitly controlling both posterior error rates when rejecting and accepting the drug respectively. While inheriting the merits of the BET design, the BETEC design is easier to implement in practice.

The rest of the paper is organized as follows. In Section 2, we present the BETEC design, and discuss its relationship with BET. We illustrate the simulation studies of the BETEC design in Section 3. Section 4 presents a trial example to assess the performance of the BETEC design. We provide a brief discussion in Section 5.

Author: Chenyang Zhang

Institute: Department of Statistics and Actuarial Science, The University of Hong Kong

Title: Bayesian nonparametric analysis for restrict mean survival time

Abstract:

Survival endpoints appear frequently in phase II and III clinical trials, and one primary focus of statistical analysis is the evaluation of treatment effect. Model-based approaches (Epstein, 1960; Cox, 1972; Bennett, 1983) have been widely used for quantifying survival benefit due to the low computational cost and desirable properties of the estimators. However, parametric estimation might be problematic and misleading if the model assumptions are violated. For example, when comparing two therapies, the hazard ratio (HR) is a common choice to assess the between-group difference under the proportional hazards (PH) assumption. If the ratio of hazard functions between two groups is not a constant over time, the estimated HR may not own a clinically meaningful interpretation (Tian et al., 2018; Yin et al., 2019). To avoid the influence of inaccurate model assumptions, nonparametric model-free estimators are proposed, such as the t -year survival rate and percentile of the survival function. However, these estimates focus mainly on local survival information and fail to provide a global summary over time.

Recently, an alternative measure called the restricted mean survival time (RMST) has attracted much research attention (Yuan and Yin, 2009; Royston and Parmar, 2013; Uno et al., 2014). The RMST is defined as the area under the survival curve up to a prespecified time τ , and can be viewed as a special case of the weighted Kaplan-Meier estimate (Pepe and Fleming, 1989) when the weight function is constant. The RMST incorporates long-term survival information free from model assumptions and provides clinically clear and meaningful interpretation as the expected survival time for patients during the follow-up period up to τ . The estimated RMST based on the Kaplan-Meier curve (Kaplan and Meier, 1958) converges to a Gaussian process (Zhao et al., 2016), for which the variance can be estimated by a perturbation-resampling method (Lin et al., 1993). The frequentist inference for the estimated RMST, e.g., the confidence interval and corresponding two-sample hypothesis testing procedure, can be easily constructed by asymptotic normal approximation, while studies on the RMST from the Bayesian nonparametric

viewpoint are limited.

In this paper, we provide a Bayesian nonparametric estimate for the posterior distribution of the RMST given right censored and interval censored observations. The Bayesian nonparametric estimation of distribution functions has been extensively studied (Ferguson, 1973; Antoniak, 1974; Susarla and Van Ryzin, 1976). We utilize the Gibbs sampler for approximating the posterior distribution of the distribution function F , which is then used for generating the posterior samples of Bayesian RMST. The proposed Bayesian RMST is shown to be a consistent and robust estimate, and can be used as a tool for Bayesian survival inference and clinical trial design.

2019

Author: Laura Harrison
Institute: Harvard TH Chang School of Public Health
Title: Power calculation for cross-sectional stepped wedge cluster randomized trials with variable cluster sizes

Abstract:

Introduction: In parallel cluster randomized trials (CRTs), ignoring variation in cluster sizes during sample size calculation leads to an under-powered study. For stepped wedge cluster randomized trials (SW-CRTs), the impact of varying cluster sizes on study power is unclear. A recent systematic review of over one-hundred SW-CRTs reported that 48% had varying cluster sizes, but only 13% accounted for this cluster size variation during sample size calculation. Standard sample size formulas for SW-CRTs assume that cluster sizes are equal.

Methods: We investigated the relative efficiency (RE) of a SW-CRT with varying cluster sizes to equal cluster sizes and derived variance estimators for the intervention effect that account for this variation under a commonly-used linear mixed effects model for cross-sectional SW-CRTs. When cluster sizes vary, the power of a SW-CRT depends on the order in which clusters receive the intervention, which is determined through randomization. We first derived a variance formula that corresponds to any particular realization of the randomization sequence and propose efficient algorithms to identify upper and lower bounds of the power. We then obtain an “expected” power based on a first-order approximation to the variance formula, where the expectation is taken with respect to all possible randomization sequences. Finally, we provide a variance formula for more general settings where only the mean and coefficient of variation (CV) of cluster sizes, instead of exact cluster sizes, are known in the design stage. A design effect and correction factor for sample size calculations that account for cluster size variation were additionally derived.

Results: We evaluated our methods through simulations and illustrated that the power of a cross-sectional SW-CRT decreases as the variation in cluster size increases, and the impact is largest when the number of clusters is small. If only the mean and CV of cluster sizes are available in the design stage, the average power can be well estimated using our methods. The efficient algorithm to identify upper and lower bounds for the power when exact cluster sizes are known gave results very close to the highest and lowest simulated powers.

Discussion: Cluster size variation should be taken into consideration in cross-sectional SW-CRT design to ensure adequate power. While the effect of unequal cluster sizes on study power seems to be smaller than for parallel CRTs; the reduction is not negligible particularly with a small number of clusters or a cluster size CV greater than one. The variance formulas we derived under a linear model are suitable for a cross-sectional design with a continuous or count outcome. In future work we aspire to investigate power and sample size formulas accounting for unequal cluster size for binary outcomes and for cohort SW-CRT designs where the same individuals are followed over time.

Author: Lee Kennedy-Shaffer
Institute: Harvard University
Title: Sample size estimation for stratified individual and cluster randomized trials with binary outcomes

Abstract:

Individual randomized trials (IRTs) and cluster randomized trials (CRTs) with binary outcomes arise in a variety of settings and are often analyzed by logistic regression and generalized estimating equations with a logit link, respectively. The effect of stratification on the required sample size is less well understood for trials with binary outcomes than for continuous outcomes. Because of this, adjusting sample size for stratification is less common when planning trials with binary outcomes. Using weighted averages of within-stratum treatment effects, we develop analytic formulae for the sample size required for stratified trials with binary outcome. We propose easy-to-use methods for sample size estimation for stratified IRTs and CRTs. These methods,

unlike previous sample size methods for stratified CRTs, work for GEEs with a logit link, do not require a common cluster size, and allow the investigator to specify any design effect. For both IRTs and CRTs, we also identify the ratio of the sample size for a stratified trial versus a comparably-powered unstratified trial, allowing investigators to evaluate how stratification will affect the required sample size when planning a trial. This requires methods to ensure comparability of within-stratum and overall treatment effects as well as within-stratum and overall design effects for CRTs. For CRTs, these methods can be used when the investigator has a priori estimates of the within-stratum intra-cluster correlations (ICCs) or, when there are no such estimates, by assuming a common within-stratum ICC. We show that this assumption is generally conservative in the two-stratum setting. Furthermore, the impact of various parameters on the effect of stratification is shown through example settings. Using these methods, we describe scenarios where stratification may have a practically important impact on the required sample size. We find that in the two-stratum case, there are unlikely to be realistically plausible scenarios in which an important sample size reduction is achieved when the overall probability of a subject experiencing the event of interest is low, both for IRTs and for CRTs with very small cluster sizes. When the probability of events is not small, or when cluster sizes are large, however, there are scenarios where practically important reductions in sample size result from stratification. We highlight scenarios where there is at least a 10% reduction in the sample size of the stratified trial compared to the unstratified trial. These results will help trial planners decide whether to stratify IRTs and CRTs and ensure that trials are appropriately sized and powered when stratification is used.

Author: Martin Law (2019 Winner)

Institute: MRC Biostatistics Unit, University of Cambridge

Title: A new class of optimally curtailed trials for phase II oncology trials

Abstract:

Most novel treatments are found to be inefficacious, which makes the average development cost associated with each successful treatment extremely high. This makes novel designs which can improve clinical research extremely valuable. Here, our focus is on achieving this within the context of single-arm phase II clinical trials with binary outcomes. Such trials generally have null hypothesis $H_0: p=p_0$. This includes Simon's design, the most frequently used phase II design amongst UK clinical trials units, and popular across the world. In this design, there is a single interim analysis, at which point stopping is allowed for a no-go decision only. Here, a no-go decision means that H_0 is not rejected and no further investigation of the treatment will take place, while a go decision would mean that H_0 is rejected and the treatment warrants further testing. Many extensions to Simon's design have been proposed, with the aim of decreasing the expected sample size: For example, allowing stopping for either a go or no-go decision when the final trial decision is certain. Ending the trial early in this manner is known as non-stochastic curtailment. A further extension is to allow stopping for either a go or no-go decision as soon as either decision becomes highly likely, known as stochastic curtailment. Designs incorporating stochastic curtailment have been proposed previously. However, these designs have allowed stochastic curtailment only when a no-go decision is likely. Further, such designs have relied on simulation to estimate trial operating characteristics, such as the expected sample size, and the search for the optimal threshold for determining when a final no-go decision is "likely" has not been comprehensive.

Here, we introduce two designs that employ stochastic curtailment for both go and no-go decisions. The exact distribution of the possible trial outcomes is calculated, meaning that the trial operating characteristics can be obtained without recourse to simulation. We search for suitable trials by undertaking a comprehensive search of thresholds for how likely a final go or no-go decision is, and further, we introduce an accurate equation for calculating this quantity, known as the conditional power, at each point in a possible trial. Moreover, rather than applying curtailment to an optimal non-curtailed design, curtailment is taken into account during the search for optimal designs. The two novel designs are compared to existing designs, across two scenarios. The designs are compared in terms of single optimality criteria, including the expected sample size. The designs are also compared using a weighted sum of optimality criteria. The best design for each possible set of weights is plotted, to give an indication of which designs perform best as optimality preferences vary. When optimising for expected sample size, the expected saving compared to Simon design ranges from 22% to 55%.

2018

Author: Kaitlyn Cook
Institute: Harvard University
Title: Futility assessment via the conditional power for cluster randomized trials with time-to-event endpoints

Abstract:

Introduction. In cluster-randomized trials (CRTs) for infectious disease prevention, time-to-event outcomes (such as time to HIV seroconversion) are often of interest. Event occurrence is assessed intermittently at pre-scheduled visits, resulting in interval-censored outcomes; cluster randomization also induces dependence between observations on individuals in the same cluster. Thus, the design, monitoring, and analysis of CRTs must account for these correlated, interval-censored data. Close interim monitoring of CRTs maximizes their chances for success by allowing for real-time study modifications. It also increases investigators' ability to assess study futility, either due to lack of efficacy or due to insufficient coverage of the intervention. Motivated by the Botswana Combination Prevention Project (BCPP), an ongoing CRT evaluating the effectiveness of a combination HIV prevention strategy in 30 communities across Botswana, we investigate conditional power-based methods for monitoring CRTs with interval-censored outcomes.

Methods. We propose a simulation-based approach to conditional power estimation. We first non-parametrically estimate the survival distributions in the intervention and control clusters based on the available interim data. We then incorporate assumptions about changes to the baseline incidence and hazard ratio over the remainder of the trial--as well as estimates of the dependency among observations in the same cluster, taken from a Cox frailty model--to project these survival curves through the end of the study. From these "full trial" curves we are able to generate correlated interval-censored observations that reasonably reflect our assumptions about the remainder of the trial. Finally, we estimate the conditional power as the proportion of times (across multiple full-data-generation steps) that the null hypothesis of no treatment effect is rejected based on a permutation test.

Results. We apply our conditional power method to a simulated interim dataset modeled on the design of the BCPP, and report conditional power estimates under a range of assumptions regarding the intervention effect over the remainder of follow-up. Simulations studies also reveal that our method provides reasonable conditional power estimates across an array of intervention effects and degrees of clustering.

Conclusion. Our simulation-based approach is a viable and flexible method for estimating the conditional power of CRTs with time-to-event endpoints.

Author: Boxian Wei (2018 Winner)
Institute: University of Michigan
Title: A Bayesian analysis of small n sequential multiple assignment, randomized trials (snSMARTs)

Abstract:

Designing clinical trials to study treatments for rare diseases is challenging because of the limited number of available patients. A suggested design is known as the small-n Sequential Multiple Assignment Randomized Trial (snSMART), in which patients are first randomized to one of multiple treatments (stage 1). Patients who respond to their initial treatment continue the same treatment for another stage, while those who fail to respond are re-randomized to one of the remaining treatments (stage 2). The data from both stages are used to compare the efficacy between treatments. Analysis approaches for snSMARTs are limited, and we propose a Bayesian approach that allows for borrowing of information across both stages. Through simulation, we compare the bias, root mean-square error (rMSE), width and coverage rate of 95% confidence/credible interval (CI) of estimators from our approach to estimators produced from (a) standard approaches that only use the data from stage 1, and (b) a log-Poisson model using data from both stages whose parameters

are estimated via generalized estimating equations. We demonstrate the rMSE and width of 95% CIs of our estimators are smaller than the other approaches in realistic settings, so that the collection and use of stage 2 data in snSMARTs provide improved inference for treatments of rare diseases.

Author: Xiaobo Zhong

Institute: Columbia University

Title: A gate-keeping test for selecting adaptive interventions under general SMART designs

Abstract:

This article proposes a method to overcome limitations in current procedures that address multiple comparisons of adaptive interventions embedded in sequential multiple assignment randomized trial (SMART) designs. Because a SMART typically consists of numerous adaptive interventions, inferential procedures based on pairwise comparisons of all adaptive interventions may suffer substantial loss in power after accounting for multiplicity. In addition, most traditional statistical methods for multiplicity adjustment in comparing non-adaptive treatments require that the correlation structure is known a priori. Since it is not the case for analyzing SMART data, these methods cannot be directly applied in SMART settings. We address these problems by proposing a likelihood-based Wald test that compares all adaptive interventions of interest in an omnibus fashion to avoid an exhaustive search, and derive its asymptotic distribution. The Wald test is applied as a gate-keeping test, which must reach a pre-specified significance level before a selection of adaptive intervention can be made, so that a false positive finding under a global null is properly controlled. We also derive the sample size calculation formula associated with the proposed test, to formally justify SMART sample sizes with respect to the pre-specified type I error rate and target power. Simulations of the proposed test show that the asymptotic approximation is accurate with a moderate sample size, and that it outperforms the existing multiple comparison procedures in terms of statistical power. Simulations also suggest that the analytical approach based on the proposed test has desirable selection properties. The application of the proposed method is illustrated with a real data set.

2017

Author: Chi Kin Lam

Institute: The University of Hong Kong

Title: Nonparametric overdose control for dose finding in drug-combination trials

Abstract:

With the emergence of novel targeted anti-cancer agents, drug combinations have been recognized as cutting-edge development in oncology. However, limited attention has been paid to the overdose control in the existing drug-combination dose-finding trials. We develop the multi-agent nonparametric overdose control (MANOC) design for dose finding in phase I drug-combination trials. Based on a Bayesian decision-theoretic approach, we control the probability of overdosing in a local region at the current dose combination. Simulation studies are conducted to investigate the performance of the proposed design. While the MANOC can prevent patients from being allocated to overtotoxic dose levels, its accuracy and efficiency are still competitive to the existing designs. As an illustration, the MANOC is applied to a phase I clinical trial for identifying the maximum tolerated dose combination of buparlisib and trametinib.

Author: Yu Lan (2017 Winner)

Institute: Southern Methodist University

Title: Adaptive prediction of event times in clinical trials

Abstract:

In event-based clinical trials it is common to plan interim analyses to take place at planned event counts. Accurate prediction of these event times can support trial planning and the efficient allocation of resources. Available methods to create such predictions include parametric cure and non-cure models and a nonparametric approach based on the Bayesian bootstrap. The parametric methods work well when their underlying assumptions are met, and the nonparametric method gives calibrated but inefficient predictions across a wide range of models. However, in the early stages of a trial, when predictions have the highest marginal value, there is insufficient data to provide evidence about the form of underlying model, including whether a cure fraction exists. In this paper, we propose an adaptive method to address this deficiency. The

method draws predictions from the model with the highest Bayesian posterior probability within a range of candidate models. To further capture the uncertainty in clinical trial prediction, we apply a simulation strategy using the Bayesian bootstrap. A simulation study demonstrates that the adaptive method produces prediction intervals that have good coverage and are slightly wider than non-adaptive intervals but narrower than nonparametric intervals. It leads to some improvements in making predictions with data from the International Chronic Granulomatous Disease Study.

Author: Ting Wang

Institute: University of North Carolina at Chapel Hill

Title: Auxiliary-variable-enriched biomarker stratified design

Abstract:

Introduction: In precision medicine, drugs are developed to target patients with certain genetic profiles. Targeted trials test treatment benefit only in the biomarker-positive patients. Trials with a biomarker-stratified design (BSD) allow a complete assessment of the effect of the new drug relative to the standard drug overall as well as in various biomarker-defined subgroups. However, a BSD trial often requires enrolling a large number of patients, especially when the proportion of the biomarker positives is small and thus the conduct of a BSD trial is expensive when the cost of ascertaining the true biomarker status is high.

Methods: We propose a special type of biomarker enrichment design, Biomarker Stratified Design Enriched by Auxiliary Variables (ABSD), in which a subgroup of patients, typically the biomarker-positive patients, are enriched based on the value of an inexpensive auxiliary variable that is positively correlated to the true biomarker. In such a design, all auxiliary-variable-positive patients and a proportion of the auxiliary-variable-negative patients are selected and included in the randomized trial. We compared the efficiency of ABSD with BSD in estimating various treatment parameters that are estimable in a BSD trial including the treatment effect in all patients and in specific biomarker subgroups and the interaction effect. We compared the efficiency of the two designs in term of the number of treated patients and the cost of the trial, assuming a range of prevalence of the true biomarker-positive patients in the overall population, the positive predictive value of the auxiliary variables for the true maker, and configurations of cost utilities of various items in conducting such trials.

Results: The proposed ABSD always reduces the total cost of the trial relative to a BSD when the prevalence rate is small and the PPV, the probability that a patient with positive auxiliary variable also has a positive true biomarker, is large enough.

When employing the proposed design in a practical study, Gefitinib or Carboplatin-Paclitaxel in Pulmonary Adenocarcinoma in North America, for testing the treatment effect among EGFR mutants and the interaction effect, ABSD requires 155 randomized patients compared to the 930 randomized patients required by a BSD. In addition, ABSD reduces the total cost cost by 64.6%.

Another advantage of ABSD is that in most cases we can immediately randomize patients selected in the screening process without waiting for the result of true biomarker test, which can substantially reduce reduce the waiting time.

Since PPV plays a very important role in the proposed design, a Bayesian adaptive ABSD is also proposed to deal with the mis-specified PPV.

Conclusion: A biomarker stratified design enriched by an auxiliary variable can be more efficient than the standard BSD design. The efficiency gain can be particularly significant when the auxiliary variable has a high PPV, the prevalence rate of the biomarker-positive subgroup is small and the cost of ascertaining the true biomarker status is high relative to the auxiliary variable.