Adaptive **D**esigns & **M**ultiple **T**esting **P**rocedures

Workshop 25th-26th April, 2024
Ibiza, Spain

Scientific Committee:
Marta Bofill Roig[1], Thomas Asendorf[2], Jordi Cortés Martínez[3], Alexandra Graf[1], Sonja Zehetmayer[1]

Organizing Committee:
Marta Bofill Roig[1], Thomas Asendorf[2], Francesc A. Rosselló Llompart[4], Irene García Mosquera[4], Arnau Mir Torres[4]

[1]Medical University of Vienna, [2]Medical University of Göttingen, [3]Universitat Politècnica de Catalunya, [4]Universitat de les Illes Balears

Sponsored by:

Organised by:

Deutsche Region

Universitat de les Illes Balears

WORKSHOP VENUE

Location:

The workshop “Adaptive Designs and Multiple Testing Procedures 2024” will take place at Hotel Vibra Algarb, located in Playa d’en Bossa, right in front of the beach.

Address:

Av. Pere Matutes Noguera, 107
07800 Eivissa, Illes Balears
Spain.

SOCIAL EVENTS

Workshop Dinner:

The workshop dinner can be attended on Thursday 25th April. Payment is the responsibility of the participant.

Guided Tour:

You are welcome to take part in a free guided tour of Ibiza's old town. Further information to come.

🌐More information at the webpage: https://admtp.github.io/ADMTP2024/index.html

SCIENTIFIC PROGRAM - OVERVIEW

Thursday, 25th April

08:00 - 08:30	Registration
08:30 - 08:40	Welcome
08:40 - 09:40	Keynote speaker I: Dominic Magirr
09:40 - 10:52	Session I: Error control
10:52 - 11:20	Coffee Break
11:20 - 12:50	Session II: Complex designs
12:50 - 14:00	Lunch break
14:00 - 15:12	Session III: Testing and estimation in group-sequential designs
15:15 - 16:15	Invited Session I: Methodological and practical outcomes from the Adaptive Designs Working Group of the MRC-NIHR Trials Methodology Research Partnership
16:15 - 16:35	Coffee break
16:35 - 17:30	Session IV: Adaptive designs
19:00	Walk and Dinner

Friday, 26th April

08:40 - 09:40	Keynote speaker II: Annette Kopp-Schneider
09:40 - 10:52	Session V: Group-sequential designs
10:52 - 11:20	Coffee Break
11:20 - 12:50	Session VI: Designs with time-to-event endpoints
12:50 - 14:00	Lunch break
14:00 - 15:12	Session VII: Multiple regression
15:15 - 16:00	WG Meeting - Coffee break
16:00 - 17:20	Invited session II: Practical experiences of using software to design clinical trials using simulations
17:20 - 17:30	Closing

SCIENTIFIC PROGRAM - DETAILED TIME SCHEDULE

Thursday, 25th April

08:00 - 08:30	Registration
08:30 - 08:40	Welcome
08:40 - 09:40	Keynote speaker I Chair: Marta Bofill Roig **Dominic Magirr**: *Deconstructing the Max-combo Test*
09:40 - 10:52	Session I: Error control Chair: Sonja Zehetmayer 1. **Vincent Jankovic**, Lasse Fischer, Werner Brannath: *Asymptotic online familywise error rate control for dependent test statistics* 2. **Ekkehard Glimm,** David Robertson: *Familywise error rate control for block response-adaptive randomization* 3. **Martin Posch**, Franz König: *Assessment of the Type 1 Error Rate after Unplanned Interim Analyses* 4. **Jelle Goeman**, Aldo Solari: *Bonferroni for Bayesians: Multiplicity Correction Based on Joint Credibility*
10:52 - 11:20	Coffee Break
11:20 - 12:50	Session II: Complex designs Chair: Carolin Herrmann 1. **Pavla Krotka**, Martin Posch, Marta Bofill Roig: *Interim analyses in platform trials with time trends and non-concurrent controls* 2. **Quynh Nguyen**, Hue Kästel, Benjamin Hofner: *A Risk to Trial Integrity – Issues of Information Leakage in Platform Trials* 3. **Sonja Zehetmayer**, Marta Bofill Roig, Martin Posch: *An adaptive basket trial design for the treatment of whorm infections* 4. **Michaela Maria Freitag**, Cornelia Ursula Kunz, Geraldine Rauch, Franz König: *Adaptive frequentist basket trials with data dependent clustering* 5. **Anastasia Ivanova**: *How to efficiently test the biomarker negative subgroup in a biomarker-stratified trial?*
12:50 - 14:00	Lunch break
14:00 - 15:12	Session III: Testing and estimation in group-sequential designs Chair: Ekkehard Glimm 1. **Pascal Rink**, Werner Brannath: *Simultaneous Bootstrap Tilting Confidence Intervals* 2. **Brice Ozenne**, Paul Blanche, Corine Baayen: *Ordering the*

	sample space for p-value and confidence interval computation in group sequential trials with delayed outcome 3. **Robin Ristl**: *A comparison of multivariate hypothesis tests for longitudinal outcomes* 4. **Abigail Jane Burdon**, Thomas Jaki: *Multivariate group sequential tests for global summary statistics*
15:15 - 16:15	Invited Session I: Methodological and practical outcomes from the Adaptive Designs Working Group of the MRC-NIHR Trials Methodology Research Partnership Chair: Thomas Asendorf 1. **Nina Wilson**: *Exploring current practices in adaptive trials: patient information sheets, costing, and efficiently conducting interim analyses* 2. **David Robertson**: *Estimation after adaptive designs* 3. **Munya Dimairo**: *Making adaptive designs more accessible. A practical adaptive designs toolkit*
16:15 - 16:35	Coffee break
16:35 - 17:30	Session IV: Adaptive designs Chair: Martin Posch 1. **Yu Shen**: *A Holistic Review of Adaptive Designs and their Regulatory Alignment* 2. **Werner Brannath**, Morten Dreher: *Lessons learned from optimal conditional error functions* 3. **Hayley Michelle Belli**, Federico Macchiavelli Giron, Andrea Troxel: *Statistical and Design Considerations for an Adaptive Stage Sequential Multiple Assignment Randomized Trial*
19:00	Walk and Dinner

Friday, 26th April

08:40 - 09:40	Keynote speaker II Chair: Thomas Asendorf **Annette Kopp-Schneider**: *Borrowing from external information in clinical trials: methods, benefits and limitations*
09:40 - 10:52	Session V: Group-sequential designs Chair: David Robertson 1. **Maria Vittoria Chiaruttini**, Giulia Lorenzoni, Dario Gregori: *The Dynamic Historical Information Borrowing in Noninferiority Bayesian Group Sequential Design for Medical Device Clinical Trials* 2. **Corine Baayen**, Paul Blanche, Brice Ozenne: *Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods* 3. **Yevgen Tymofyeyev**, Michael Grayling: *Automation tools for group sequential designs with MTP and SSR* 4. **Samuel Sarkodie**, James Wason, Michael Grayling: *Optimal drop-the-loser trials when an intermediate endpoint is used for interim selection*
10:52 - 11:20	Coffee Break
11:20 - 12:50	Session VI: Designs with time-to-event endpoints Chair: Robin Ristl 1. **Carolin Herrmann**, Paul Blanche: *Sample size adaptations under non-proportional hazards using the restricted mean survival time* 2. **Jan Meis**, Carolin Herrmann, Björn Bokelmann, Meinhard Kieser: *Blinded-into-unblinded interim analyses for time-to-event-trials with fixed analysis timings: a simulation study* 3. **Merle Munko**, Marc Ditzhaus, Dennis Dobler, Jon Genuneit: *Surviving the multiple testing problem: RMST-based tests in general factorial designs* 4. **Moritz Fabian Danzer**, Ina Dormuth: *Interim adaptations of weights in survival testing procedures* 5. **Jordi Cortés Martínez**, Marta Bofill Roig, Guadalupe Gómez Melis: *CompAREdesign: An R Package for the Design of Randomized Clinical Trials with Composite Endpoints*
12:50 - 14:00	Lunch break
14:00 - 15:12	Session VII: Multiple regression Chair: Anastasia Ivanova 1. **Beatrijs Moerkerke**, Tom Loeys, Kelly Van Lancker: *Assessing the role of interim analyses in addressing*

	replication concerns in behavioral sciences 2. **Lara Vankelecom**, Ole Schacht, Tom Loeys, Beatrijs Moerkerke: *Sample Size Re-Estimation as an Alternative to A Priori Power Analysis in Social Sciences: The Multiple Linear Regression Case* 3. **Ole Schacht**, Lara Vankelecom, Tom Loeys, Beatrijs Moerkerke: *Information-Based Monitoring as an Alternative to A Priori Power Analysis in Social Sciences: The Multiple Linear Regression Case* 4. **Thorsten Dickhaus**, Vladimir Vutov: *Multiple marginal models for multinomial regression with high-dimensional covariates*
15:15 - 16:00	WG Meeting - Coffee break
16:00 - 17:20	Invited session II: Practical experiences of using software to design clinical trials using simulations Chair: Elias Laurin Meyer 1. **Peter Jacko**: *Using the "SIMulating PLatform trials Efficiently" (SIMPLE) R package to develop a simulator for a bespoke platform trial* 2. **Gernot Wassmer & Friedrich Pahlke**: *Flexible Clinical Trial Planning with the R Package rpact* 3. **Tobias Mielke**: *Design or simulation: What comes first?* **Discussant: Daniel Sabanés Bové**
17:20 - 17:30	Closing

ABSTRACTS

Keynote speaker I

Thursday, 25th April, 08:40 - 09:40

Deconstructing the Max-combo Test

Dominic Magirr (Novartis Pharma AG)

The Max-Combo Test (Lin et al., 2020) is an adaptive procedure to select the best test statistic from a small prespecified set of candidates. It includes a correction for multiplicity. Proposed applications include clinical trials in oncology when there is prior uncertainty regarding the commonly made proportional hazards assumption. In this context, the candidate set of test statistics usually come from the Fleming-Harrington rho-gamma family of weighted log-rank statistics. In some of the candidate test statistics, the weighting is tilted towards events occurring in the early part of follow up, whereas in others it is tilted towards events in the later part of follow up. The overall procedure is therefore robust to a wide range of treatment effect patterns.

In this talk I shall take a close look at the Max-Combo Test, using recently proposed visualization techniques (Jimenez et al., 2023) to explain some of its counter-intuitive properties, as pointed out in recent publications (Freidlin & Korn, 2019). The same visualization techniques can be used to re-evaluate weighted log-rank tests in the broader context of the estimand framework in clinical trials.

References

- Lin, R.S., Lin, J., Roychoudhury, S., Anderson, K.M., Hu, T., Huang, B., Leon, L.F., Liao, J.J., Liu, R., Luo, X. and Mukhopadhyay, P., 2020. Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. Statistics in Biopharmaceutical Research, 12(2), pp.187-198
- Fleming, T.R. and Harrington, D.P., 1981. A class of hypothesis tests for one and two sample censored survival data. Communications in Statistics-Theory and Methods, 10(8), pp.763-794.
- Jiménez, J.L., Barrott, I., Gasperoni, F. and Magirr, D., 2023. Visualizing hypothesis tests in survival analysis under anticipated delayed effects. arXiv preprint arXiv:2304.08087.
- Freidlin, B. and Korn, E.L., 2019. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis?. Journal of Clinical Oncology, 37(35), p.3455.

Session I: Error control

Thursday, 25th April, 09:40 - 10:52

Asymptotic online familywise error rate control for dependent test statistics

Vincent Jankovic, Lasse Fischer, Werner Brannath
University of Bremen, Germany; jankovic@uni-bremen.de

In online multiple testing, an a priori unknown number of hypotheses are tested sequentially, i.e. at each time point a test decision for the current hypothesis has to be made using only the data available so far. Although many powerful test procedures have been developed for online error control in recent years, most of them are designed solely for independent or at most locally dependent test statistics. In this work, we provide a new framework for deriving online multiple test procedures which ensure asymptotical (with respect to the sample size) control of the familywise error rate (FWER), regardless of the dependence structure between test statistics. In this context, we give a few concrete examples of such test procedures and discuss their properties. Furthermore, we conduct a simulation study in which the type I error control of these test procedures is also confirmed for a finite sample size and a gain in power is indicated.

Familywise error rate control for block response-adaptive randomization

Ekkehard Glimm[1], David Robertson[2]
[1]Novartis Pharma, Switzerland; [2]MRC Biostatistics, University of Cambridge, UK;
ekkehard.glimm@novartis.com

Recent years have seen several initiatives aiming at the advancement of complex innovative designs. EU-PEARL (EU-PEARL - Innovative Patient Centric Clinical Trial Platforms) and the FDA's Complex Innovative Trial Design Meeting Program | FDA are examples.
These initiatives have brought about a renewed interest in response-adaptive randomization. In clinical studies with response-adaptive randomization, response data collected in the study is used to modify the randomization algorithm. The technique is commonly used for studies in the early phases of clinical development. For confirmatory studies, however, it is rarely applied. One of the reasons for this is that type I error control poses challenges for this type of design.
In this paper, we will present a method for type I error control in response-adaptive trials (Glimm and Robertson, 2023). Adaptations are done following the response from blocks of patients. It is a modification of a more flexible, but less powerful approach suggested by Robertson and Wason (2019). The techniques we are using are similar to the recursive combination tests suggested by Brannath et al. (2002).
The talk will outline the technique, present an example of its application and compare its power with that of alternative approaches that do not control the type I error rate.

References
- Glimm E and Robertson, DS: Familywise error rate control for block response-adaptive randomization. Statistical Methods in Medical Research 2023; 32: 1193–1202.
- Brannath W, Posch M and Bauer P. Recursive combination tests. J Am Stat Assoc 2002; 97: 236–244.
- Robertson DS, Wason J. Familywise error control in multi-armed response-adaptive trials. Biometrics 2019; 75: 885–894.

Assessment of the Type 1 Error Rate after Unplanned Interim Analyses

Martin Posch, Franz König
Medical University of Vienna, Austria; martin.posch@meduniwien.ac.at

Unplanned interim analyses are sometimes necessary in clinical trials but can lead to an inflation of the Type 1 error rate. The extent of the inflation depends on several factors, including the type of information that triggered the interim analysis (such as external data, or blinded or unblinded data of primary, secondary or safety endpoints), the process used to trigger the analysis (continuous monitoring or a single look), the potential implications of the unplanned interim analysis (early rejection, futility stopping, sample size reassessment, treatment or endpoint selection), and the statistical methods employed. To address the issue of Type 1 error rate inflation, adjustments can be made using the conditional error rate principle. However, this method requires pre-specification of the maximum information and may be challenging, e.g., in trials with time-to-event endpoints if the time of the primary analysis is not event driven. In this work we explore in several examples how the type 1 error rate after unplanned interim analysis can be assessed. Especially, we consider different scenarios with regard to the triggers of the interim analysis, type of adaptation and analysis method.

Bonferroni for Bayesians: Multiplicity Correction Based on Joint Credibility

Jelle Goeman[1], Aldo Solari[2]
[1]Leiden University Medical Center, The Netherlands; [2]Ca' Foscari University of Venice, Italy; j.j.goeman@lumc.nl

It is commonly argued that correction for multiplicity is not needed in the Bayesian paradigm. If multiplicity is addressed at all in Bayesian analyses, it is often dealt with by choosing an appropriate prior. In this paper we look at the issue of multiplicity of inference in Bayesian analysis from the alternative perspective of the posterior, and its summary in terms of credible intervals. From this perspective, we construct a rationale for multiplicity adjustment, adopting a fully subjectivist Bayesian point of view. Within this framework, emphasizing joint credibility, direct Bayesian equivalents of the methods of Bonferroni, Min-p, and Scheffé emerge naturally. Our findings apply directly to inference based on credible intervals, but

may also be generalized to Bayesian hypothesis testing when such testing is based on posterior probabilities.

Session II: Complex designs

Thursday, 25th April, 11:20 - 12:50

Interim analyses in platform trials with time trends and non-concurrent controls

Pavla Krotka, Martin Posch, Marta Bofill Roig
Medical University of Vienna, Austria; pavla.krotka@meduniwien.ac.at

Platform trials accelerate drug development by offering increased flexibility and efficiency. They evaluate the efficacy of multiple treatment arms under a single master protocol, with the added benefit of permitting treatment arms to enter the trial over time and to stop early based on interim data. Treatment efficacy is usually assessed using a shared control arm. For arms entering later, the control data is divided into concurrent and non-concurrent controls (NCC), referring to control patients recruited while the given treatment arm is in the platform and before it enters, respectively. Analysis using NCC can reduce the required sample size and increase power, but might also lead to bias in the effect estimates and hypotheses tests, if there are time trends.
For platform trials with continuous endpoints without interim analyses, a regression model has been proposed that utilizes NCC and adjusts for time trends by including the factor "period" as a fixed effect. Here, periods are defined as time intervals bounded by any treatment arm entering or leaving the platform. It was shown that this model leads to unbiased effect estimates and asymptotically controls the type I error rate regardless of the time trend pattern, if the time trend affects all arms in the trial equally and is additive on the model scale. However, so far it has not been explored if this method also leads to valid testing and estimation procedures in platform trials with interim analyses that allow for early stopping.
In this talk, we will explore the properties of various regression models that adjust for time trends in platform trials using NCC data. Especially, we will investigate different variants of the time adjustment, considering splitting the trial into periods or calendar time intervals of fixed length that are not dependent on the trial design. We will present results from a simulation study, evaluating the performance of the considered approaches in terms of the type I error rate and statistical power for individual treatment-control comparisons under a wide range of settings. In particular, we will consider platform trials without interim analyses, as well as trials that include interim looks.

A Risk to Trial Integrity – Issues of Information Leakage in Platform Trials

Quynh Nguyen, Hue Kästel, Benjamin Hofner
Section Data Science & Methods, Paul-Ehrlich-Institut, Germany;
QuynhLan.Nguyen@pei.de

Platform trials offer an opportunity to study multiple treatments at the same time using a shared control, and allow further treatments to be added to an ongoing trial. While this design provides flexibility, concerns about trial integrity and maintaining the blind arise due to the increased complexity.
Within the framework of the EU clinical trial regulation, timely submission or publication of summary results is mandated within one year of completion, or even earlier if stock market requirements apply. The comparative analysis results for one treatment against a shared control, such as the mean and 95% confidence intervals or median survival times, can inadvertently reveal information about the shared control, potentially compromising the blind for the remaining treatments on the trial. Supplemented with access to information from regular data monitoring activities, such as pooled data from efficacy endpoints for data cleaning or trial adaptations, or simply the monitoring of the required number of events for event-driven analyses, allows for the recalculation or estimation of current comparative results for treatments still on the trial.
We present one possibility for an event-driven trial with a common control where the release of the hazard ratio and median survival times of one treatment arm can provide enough information for the estimation of the current hazard ratio for another treatment still on the trial if certain assumptions are made for the survival endpoint. Based on the estimated recalculated hazard ratio, a variety of trial adaptations can be performed. However, since these changes to the trial design are to some extent data-driven, trial integrity might be at risk. To address this concern, we present additional measures that should be implemented to prevent or mitigate potential risks to trial integrity.

An adaptive basket trial design for the treatment of whorm infections

Sonja Zehetmayer, Marta Bofill Roig, Martin Posch
Meduniwien, Austria; sonja.zehetmayer@meduniwien.ac.at

We propose a frequentist, adaptive basket trial design to investigate the safety and efficacy of three different doses compared to placebo to treat four types of worm infections (corresponding to 4 baskets). As the safety of the highest dose is not yet established, the study starts with the two lower doses (and control). Then, based on safety and efficacy results observed in an interim analysis, it is decided to either continue with the two lower doses or to drop one or both of these doses and to start an arm with the highest dose instead. This basket trial design borrows information across baskets for the safety assessment but efficacy is assessed for each basket separately. However, the trial includes co-infected patients who are member of more than one basket and are used in an exploratory analysis.
The proposed adaptive design addresses three additional challenges (1) the primary endpoint is measured 12 month after recruitment and is not observed at the interim analysis.

Therefore, the adaptation decisions have to rely on an early endpoint. (2) The primary outcome variable has a mixture distribution with a lognormal component and an atom at 0. (3) the sample size per arm, should be reassessed, depending on the number of arms investigated in the second stage, such that the overall sample size remains fixed.
To control the familywise error rate (adjusting for the comparison of multiple doses to a control) in the adaptive design the partial conditional error approach is extended to allow for the inclusion of new hypotheses after an interim analysis. In a comprehensive simulation study a range of design options and analysis strategies were compared and the robustness of the design with respect to design assumptions and parameter values was investigated.
The simulation results demonstrate under which conditions the adaptive design enhances the trial's efficiency to identify the optimal dose. Adaptive dose selection allows for resource allocation to promising treatment arms and thereby can increase the chance to select the optimal dose while reducing the required overall sample size and trial duration.

Adaptive frequentist basket trials with data dependent clustering

Michaela Maria Freitag[1], Cornelia Ursula Kunz[2], Geraldine Rauch[1,3], Franz König[4]
[1]Charité - Universitätsmedizin Berlin, Germany; [2]Böhringer Ingelheim; [3]Technische Universität Berlin; [4]Medizinische Universität Wien; michaela-maria.freitag@charite.de

Traditionally, cancer treatments were based on the location of the primary tumor. However, advances in precision medicine shifted this paradigm towards molecular-targeted drugs, which can be efficacious in multiple tumor sides. By this, treatments become more patient-individualized, resulting in novel statistical challenges like stratification, small sample sizes, and adaptive decision making. One way of incorporating molecular profiling into drug development is using basket designs. In oncologic basket trials, a single drug is simultaneously tested in various tumor locations, where each location defines a substudy, the so-called “basket”. These baskets share a common study protocol, streamlining logistic, legal, clinical, and statistical aspects as much as possible. Besides the common infrastructure, the gain in efficiency is partly due to the exchange of data between the substudies.
While numerous Bayesian designs with borrowing across baskets have been proposed and compared, substantially fewer frequentist designs were developed and methodological comparisons between those are still lacking.
We compare, develop and improve frequentist approaches addressing clustering in basket trials, including the very recently published methods by Hattori and Morita (2023), and Kanapka and Ivanova (2023).
We evaluate these methods using various performance indicators, including marginal power, familywise error rate, and correctness of clusters, to provide recommendations on the best-suited method for specific applications. Furthermore, we modify these methods by e.g. incorporating optimal futility stopping rules in the clustering process.
This comprehensive analysis aims to advance the understanding and application of frequentist approaches in oncologic basket trials, ultimately contributing to more effective and individualized cancer treatments.

How to efficiently test the biomarker negative subgroup in a biomarker-stratified trial?

Anastasia Ivanova
UNC at Chapel Hill, USA, United States of America; aivanova@bios.unc.edu

In a clinical trial with a predefined subgroup, it is assumed that the biomarker positive subgroup has the same or higher treatment effect compared to its complement, the biomarker negative subgroup. In registration trials, the treatment effect is usually evaluated in the biomarker positive subgroup and in the whole population. Statistical testing of the treatment effect in the biomarker negative subgroup is usually not done since it requires a larger sample size. As a result, the new intervention can be shown effective in the overall population even though it is only effective in the biomarker positive group. To improve decision making in such trials, one needs to test the treatment effect in the biomarker negative subgroup as well as the biomarker positive subgroup. We propose an efficient way to do that.

Session III: Testing and estimation in group-sequential designs

Thursday, 25th April, 14:00 - 15:12

Simultaneous Bootstrap Tilting Confidence Intervals

Pascal Rink, Werner Brannath
Institute for Statistics and Competence Center for Clinical Trials, University of Bremen, Germany; p.rink@uni-bremen.de

Bootstrap tilting is a universal technique to improve the accuracy of statistical parameter estimation and confidence interval construction. It is particularly useful when dealing with non-normally distributed data. Bootstrap tilting combines the generation of bootstrap samples from the data at hand and the application of a transformation on each of these bootstrap samples, accounting for distributional characteristics of the data. This transformation is designed such that the transformed bootstrap sample resembles a specific target distribution, for example, a specific null distribution, which is helpful when estimating a confidence interval.
In this talk, I will present how to extend bootstrap tilting confidence intervals to a multiple testing setup. In particular, we propose to introduce a maxT-type multiplicity correction to the estimation in order to obtain simultaneous confidence intervals for multiple statistical parameters. The resulting statistical inference is valid, as the proposed simultaneous

confidence intervals are consistent for mean-type distributional parameters. I will also present simulation results that illustrate how our method compares to standard approaches in finite samples, and an application to real-world data.

Ordering the sample space for p-value and confidence interval computation in group sequential trials with delayed outcome

Brice Ozenne[1,2], Paul Blanche[1], Corine Baayen[3,4]
[1]University of Copenhagen, Section of Biostatistics, Denmark; [2]Neurobiology Research Unit, Rigshospitalet, Copenhagen; [3]Biometric Division, H. Lundbeck A/S, Valby, Denmark; [4]Global Biometrics, Ferring Pharmaceuticals, Copenhagen, Denmark; brice.ozenne@nru.dk

In group sequential trials, recruitment may be stopped at pre-defined interim analyses for efficacy or futility. Compared to trials with a fixed sample size, the testing procedure does not only rely on the value of the test statistic upon stopping but also on at which interim the trial was stopped. The p-value, probability under no treatment effect to obtain a more extreme result than the observed one, is thus defined relative to a two dimensional space. Favoring rejections at early interim analyses (regardless of the test statistics) and rejections with larger test statistics when at the same interim analyses, correspond to the Armitage ordering. In this talk, we will discuss the adaptation of this ordering to the case of a specific group sequential trial design for delayed outcome introduced by Hampson & Jennison (2013). Indeed, when the trial is stopped at an interim analysis, recently recruited patients will typically have incomplete follow-up. A decision analysis should then be performed when data on all recruited patients have been collected to conclude about treatment efficacy. We will show how to include these extra analyses in the Armitage ordering and illustrate the p-value space, i.e., how p-values vary across decision analyses and test statistic values with this new ordering. Further extension of the ordering to handle non-binding futility rules and/or constraints on the efficacy boundaries to exceed 1.96 can be achieved via conservative approaches. Considering the p-value space and results of simulation studies, we will discuss why conservative approaches can be problematic for computing the confidence interval. We will refer (but not present) to our R package DelayedGSD, freely available on Github, for the implementation of the proposed methods to compute p-values and confidence intervals.

A comparison of multivariate hypothesis tests for longitudinal outcomes

Robin Ristl
Medical University of Vienna, Austria; robin.ristl@meduniwien.ac.at

In many randomized clinical trials, a metric outcome is measured repeatedly at subsequent pre-scheduled time-points and the aim of the study is to establish a difference in the longitudinal outcome. The trialist needs to decide whether to test a null hypothesis of no difference at a predefined single time-point or a multivariate null hypothesis of no difference at any time-point. The latter approach, when applied to clinically relevant time-points, may be preferred for reasons of robustness if the timing of a treatment effect is not well known in

advance, and for reasons of efficiency if a treatment effect at more than one time-point is expected. A standard approach is to fit a mixed model for repeated measures and test a treatment by time interaction via an ANOVA-type F- or Chi-Squared test. However, depending on the correlation between repeated measures, these tests may exhibit a non-monotonic power function such that the power can decrease if the treatment effect at a specific time-point is increased. Alternative methods include a test for the maximum difference, a test for the maximum standardized difference and a test for the difference in areas under the curves.
Motivated from two studies in oncology and neurology with longitudinal outcomes, we systematically assessed the impact of different covariance structures, mean difference patterns and number of time-points on the operating characteristics of the beforementioned tests. Power calculations were performed via simulation and via numeric integration under an asymptotic normal approximation. We considered tests for a global null hypthesis of no treatment effect, as well as closed testing procedures based on the different hypothesis tests.
Our results provide systematic guidance on the choice of an appropriate hypothesis test and on the respective power and sample size calculations.

Multivariate group sequential tests for global summary statistics

Abigail Jane Burdon[1], Thomas Jaki[1,2]
[1]University of Cambridge, United Kingdom; [2]University of Regensburg, Germany;
abigail.burdon@mrc-bsu.cam.ac.uk

We describe group sequential tests which efficiently incorporate information from multiple endpoints allowing for early stopping at pre-planned interim analyses. We formulate a testing procedure where several outcomes are examined, and interim decisions are based on a global summary statistic. An error spending approach to this problem is defined which allows for unpredictable group sizes and nuisance parameters such as the correlation between endpoints. We present and compare three methods for implementation of the testing procedure including numerical integration, the Delta approximation and Monte Carlo simulation. In our evaluation, numerical integration techniques performed best for implementation with error rate calculations accurate to five decimal places. Our proposed testing method is flexible and accommodates general, non-linear, summary statistics informed by the statistical model. Type 1 error rates are controlled, and sample size calculations can easily be performed to satisfy power requirements.

Invited session I: Methodological and practical outcomes from the Adaptive Designs Working Group of the MRC-NIHR Trials Methodology Research Partnership

Thursday, 25th April, 15:15 - 16:15

Exploring current practices in adaptive trials: costing and efficiently conducting interim analyses

Nina Wilson
Newcastle University, United Kingdom; nina.wilson@newcastle.ac.uk

Adaptive designs are increasingly being recognised as an important part of improving the efficiency of clinical trials. Nevertheless, some barriers to their use remain. One important barrier is lack of understanding about the additional resources required to conduct a high-quality adaptive clinical trial, compared to a more standard trial. Secondly, is the difficulty in performing the potential interim analyses in a timely manner while maintaining quality. This talk will discuss the Costing Adaptive Trials (CAT) project and ROBust Interim analyses for adaptive trials (ROBIN) project, both funded by NIHR in the United Kingdom.
CAT was set up to:
1) investigate the additional costs that result from designing, conducting and analysing adaptive trials;
2) provide guidance on what additional costs should be included in future applications;
3) identify methodology needs for reducing the additional costs of conducting adaptive trials in the future.
ROBIN aimed to:
1) investigate what are the best practice approaches for conducting interim analyses in a quick and high-quality way;
2) look for methods, procedures or tools that can help make interim analyses quicker and higher-quality to ensure adaptive designs provide maximum benefit.
This presentation will: shed light on the additional costs required to adequately support a high-quality adaptive trial; provide guidance on how to appropriately resource an adaptive trial; and provide advice on how to conduct efficient interim analyses. Together, these projects will help ensure adaptive designs are not being prevented from their potential to add efficiency to clinical research.

Estimation after adaptive designs

David Robertson
MRC Biostatistics Unit, University Of Cambridge, United Kingdom;
david.robertson@mrc-bsu.cam.ac.uk

The use of adaptive designs for clinical trials can bring substantial benefits in terms of efficiency and patient benefit. However, while the hypothesis testing framework for adaptive designs is well-established, the best way to estimate treatment effects in such trials remains an open question. This talk aims to provide an overview of the statistical challenges associated with estimating treatment effects in adaptive clinical trials and to provide practical guidance on how to choose appropriate statistical methods and report trial results. More specifically, we will describe the challenges and limitations of using standard point estimators and confidence intervals to estimate treatment effects and provide an overview of the different types of statistical methods for estimation in adaptive designs. We will also provide practical considerations and guidance around how to choose a method and report trial results, as illustrated using a case study. This talk is based on recent and ongoing work of the UK Adaptive Designs Working Group of the MRC-NIHR Trials Methodology Research Partnership.

Making adaptive designs more accessible: a practical adaptive designs toolkit

Munya Dimairo[1], Mike Bradburn[1], Laura Flight[2], Thomas Jaki[3], Philip Pallmann[4], Graham M Wheeler[5], Cindy Cooper[1]
[1]Clinical Trials Research Unit (CTRU), Sheffield Centre for Health and Related Research (SCHARR), University of Sheffield; [2]Health Economics and Decision Science (HEDS), SCHARR, University of Sheffield; [3]MRC Biostatistics Unit, University of Cambridge; Faculty of Informatics and Data Science, University of Regensburg; [4]Centre for Trials Research, Cardiff University; [5]Imperial Clinical Trials Unit, Imperial College London; GSK;
m.dimairo@sheffield.ac.uk

Background
Adaptive designs (ADs) can help improve the way we conduct clinical trials. However, the lack of practical knowledge in ADs among diverse clinical trial stakeholders hinders their routine use, although it is steadily improving. To help address this, we developed an online, open-access, comprehensive, flexible, and practical educational toolkit on ADs for clinical trial stakeholders including non-statisticians.
Methods
We iteratively developed practical educational material covering several different topics relating to ADs. This was informed by prior work and diverse practical knowledge and experience of the project team in ADs and non-adaptive trial designs. The Practical Adaptive and Novel Designs & Analysis (PANDA) toolkit was developed using “Ruby on Rails” with a React frontend and is hosted on the University of Sheffield servers. We sought feedback

from users on the toolkit design, webpage content, and structure throughout the project. Further feedback is continuously collected from users via email (panda.sheffield.ac.uk) to improve the toolkit.

Results

The PANDA toolkit is available at https://panda.shef.ac.uk. PANDA allows self-paced practical learning that is easily accessible to anyone involved in clinical trials research. PANDA users can learn remotely about ADs at a time that suits them, and easily find content relevant to them focused on different stages of a trial with the aid of the search function. Topics covered include:

- lay description of an AD, the goals of different types of ADs, and research questions they can help address;
- potential benefits and limitations of different types of ADs;
- advice on communicating ADs to key stakeholders;
- how to cost an adaptive trial in a grant application;
- statistical methods underpinning different ADs, focusing on the design, monitoring, and analysis;
- case studies illustrating the design (with Stata and R code), communication, monitoring and analysis;
- reporting guidance;
- considerations for health economics evaluations in adaptive trials;
- measures to minimise operational and statistical biases when running an adaptive trial.

This talk will cover the functionalities within the PANDA platform.

Conclusions

The PANDA toolkit is a globally accessible resource that will evolve in response to research needs and feedback from users. We hope it will be a vital educational resource to increase practical knowledge and appropriate uptake of adaptive trials for years to come, improving clinical trial efficiency.

Session IV: Adaptive designs

Thursday, 25th April, 16:35 - 17:30

A Holistic Review of Adaptive Designs and their Regulatory Alignment

Yu Shen
UT MD Anderson Cancer Center, United States of America; yshen@mdanderson.org

When conducting a large-scale clinical trial, it is crucial to sequentially monitor the trial for ethical, scientific, and economic considerations. Utilizing cumulated data from interim analyses to adaptively make decisions during the trial's progression represents a dynamic

process. In this presentation, I will offer a comprehensive review of adaptive clinical trial designs within the frequentist framework and explore their connection with Bayesian methods. Our team has developed an adaptive trial design aimed at modifying the maximum sample size based on accumulating data in the ongoing trial without inflating the overall false positive rate. We have extended this trial design from continuous outcomes to right-censored outcomes, ensuring unbiased estimation of the primary parameter and its exact confidence interval while maintaining the correct nominal level in a sequential adaptive clinical trial. It is imperative that the proposed adaptive designs align with regulatory standards while providing flexibility to attain sufficient statistical power under true alternative hypotheses and adjust the sample size appropriately. Computations and applications will be discussed.

Lessons learned from optimal conditional error functions

Werner Brannath, Morten Dreher
University of Bremen, Germany; brannath@uni-bremen.de

The efficiency of adaptive designs has been discussed from the very beginning and is still an issue today. Accordingly, the development of optimal adaptive designs is an important task that was addressed early on and is still addressed today. In this talk, we will review a now 20-year-old optimality theory for adaptive designs in which the sample size is recalculated for a given conditional power. This theory has the advantage of providing an analytical expression for the conditional error function that minimises the expected sample size under a given parameter value or a predetermined mixture of parameter constellations. Moreover, it provides a kind of maximum likelihood approach and allows optimisation under constraints, such as a minimum and maximum for the sample size. We will present lessons to be learnt from this theory and discuss recent extensions.

References

- Brannath, W., Bauer P. (2004). Optimal Conditional Error Functions for the Control of Conditional Power. Biometrics 60, 715–723.
- Brannath, W., Dreher, M. (2024). Optimal monotone conditional error functions. arXiv:2402.00814

Statistical and Design Considerations for an Adaptive Stage Sequential Multiple Assignment Randomized Trial

Hayley Michelle Belli, Federico Macchiavelli Giron, Andrea Troxel
New York University, United States of America; Hayley.Belli@nyulangone.org

In the precision medicine era, there is a need to design patient-focused, pragmatic clinical trials. In this talk, we will introduce an approach for determining individualized treatment duration in a Sequential Multiple Assignment Randomized Trial (SMART). SMARTs are an adaptive design where every participant is first randomized to a treatment or control arm, similar to a classic parallel design. However, following this initial assignment, patients move

through a series of stages with the option to be re-randomized to switch treatments, depending on their response to the intervention in the stage prior. The SMART framework mimics standard clinical practice in that with time patients will have the opportunity to be assigned to more effective treatments all within the rigorous experimental framework of a randomized controlled trial. However, a limitation to the SMART design is that the duration of the stages is applied uniformly to all participants and is selected by investigators in advance. Since the primary objective of this design is to arrive at an optimal set of decision rules, the duration for which to administer a treatment is an important component of the dynamic treatment regime and should be derived experimentally, rather than selected a priori.
In the present work, we introduce the concept of an adaptive stage SMART design. We propose an algorithm that uses a likelihood-based approach to determine when a patient should stay on a treatment for the full stage duration or switch interventions prior to the stage end. We first derive the algorithm, and then demonstrate its performance using data from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) Study, a two-stage SMART design with multiple interim data points measuring the effectiveness of sertraline in 242 patients with nonpsychotic Major Depressive Disorder (Trivedi et al. 2016. Journal of Psychiatric Research, 78: 11-23). We discuss results from simulations that explore the use of limited interim data (only two or three measurements within a single stage) to determine whether and when a patient should be re-randomized. By varying the frequency and timing of these data, simulations reveal true positive rates between 70-90% and false positive rates between 20-50%. We end by discussing some practical considerations for analyzing data, implementation, and testing intervention effectiveness within the proposed adaptive stage SMART design framework.

Keynote speaker II

Friday, 26th April, 08:40 - 09:40

Borrowing from external information in clinical trials: methods, benefits and limitations

Annette Kopp-Schneider (German Cancer Research Center)

When trials can only be performed with small sample sizes as, for example, in the situation of precision medicine where patients cohorts are defined by a specific combination of biomarker and targeted therapy, borrowing information from historical data is currently discussed as an approach to improve the efficiency of the trial. In this context, borrowing information is often also referred to as evidence synthesis or extrapolation, where external data could be historical data or another source of co-data. A number of approaches for borrowing from external data that dynamically discount the amount of information transferred from external data based on the discrepancy between the external and current data have been proposed. We will present two selected approaches. The robust mixture prior (Schmidli et al, 2014) is a popular method. It is a weighted mixture of an informative and a robust prior, equivalent to a meta-analytic-combined analysis of historical and new data, assuming that parameters are exchangeable across trials. The power prior approach incorporates external data in the prior used for analysis of the current data. This prior is proportional to the likelihood of the external data raised to the power of a weight parameter. An Empirical Bayes approach for the estimation of the weight parameter from the similarity of external and current data has been proposed by Gravestock et al. (2017).

We will discuss the frequentist operating characteristics (FOC) of trials using these two adaptive borrowing approaches, evaluating type I error rate and power as well as Mean Squared Error. Use of the robust mixture prior requires the selection of the mixture weight, the mean and the variance of the robust component and we will discuss the impact of the selection on FOC. The concept of prior effective sample size facilitates quantification and communication of prior information by equating it to a sample size. When prior information arises from historical observations, the traditional approach identifies the ESS with a historical sample size, a measure that is independent of the current observed data, and thus does not capture an actual loss of information induced by the prior in case of prior-data conflict. The effective current sample size of a prior (Wiesenfarth and Calderazzo 2020) is introduced which relates prior impact to the number of (virtual) samples from the current data model. All aspects that will be discussed show that in the frequentist perspective borrowing cannot be beneficial for any possible true parameter value. However, benefits can be obtained if prior information is reliable and consistent.

References

- Gravestock I, Held L (2017). Adaptive power priors with empirical Bayes for clinical trials. Pharmaceutical Statistics 16:349-360.

- Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics, 70(4), 1023-1032.
- Wiesenfarth M, Calderazzo S (2020). Quantification of prior impact in terms of effective current sample size. Biometrics 76(1), 326-336.

Session V: Group-sequential designs

Friday, 26th April, 09:40 - 10:52

The Dynamic Historical Information Borrowing in Noninferiority Bayesian Group Sequential Design for Medical Device Clinical Trials

Maria Vittoria Chiaruttini, Giulia Lorenzoni, Dario Gregori
Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic and Vascular Sciences, University of Padova, Padova, Italy, Italy;
mariavittoria.chiaruttini@ubep.unipd.it

The field of medical device technology evolves rapidly with shorter lifecycles compared to pharmaceuticals. This acceleration emphasizes the need for swift safety and efficacy assessments to match technological advancements. In producing new evidence, leveraging clinical data from previous device versions is crucial, especially as medical devices are often built upon similar mechanisms.
In this context, Bayesian statistics represents a valuable approach since it naturally combines prior information with current information on a quantity of interest. Mainly, Bayesian Dynamic Borrowing (BDB) has emerged as an approach that can adjust the weight of historical information as current data accumulates, depending on the congruence between past and new data allowing for unbiased data augmentation.
Moreover, at the design stage, the synergy between BDB and group sequential design (GSD) can improve the operating characteristics of the trial and reduce the sample size.
The present study delves into BDB employing two innovative methods—Normalized Power Prior (NPP) and Self-Adapting Mixture (SAM) prior—in conjunction with the GSD theory within a noninferiority trial design to assess the operating characteristics under both congruence and incongruence scenarios between past and current data. A cardiovascular medical device trial serves as a motivating example, showcasing the practical application of these methods.
Simulation studies conducted under congruence scenarios underscore the efficiency of the GSD when combined with both NPP and SAM methods, resulting in a reduction in sample size while maintaining Type I error and Power at nominal levels thanks to the application of alpha and beta spending functions to deal with interim analyses. When exploring incongruent scenarios, the NPP design surpasses the SAM prior, exhibiting enhanced

control over Type I error and consistently leading to a more substantial reduction in sample size.
In conclusion, this paper emphasizes the potential advantages of incorporating BDB, especially when integrated with GSD, in the rapidly evolving landscape of the medical device development chain. The analyzed scenarios underscore the critical importance of carefully selecting the most suitable approach for specific trial hypotheses and the GSD boundaries with the aid of simulations.

Design and analysis of group sequential trials for repeated measurements when pipeline data occurs: a comparison of methods

Corine Baayen[1,2], Paul Blanche[3], Brice Ozenne[3,4]
[1]Global Biometrics, Ferring Pharmaceuticals, Denmark; [2]Biometrics Division, H. Lundbeck A/S, Denmark; [3]Department of Public Health, Section of Biostatistics, University of Copenhagen, Denmark; [4]Neurobiology Research Unit and Brain Drugs, Copenhagen University Hospital, Denmark; Corine.Baayen@ferring.com

Group sequential trials allow for early stopping of a clinical trial for efficacy or futility, without compromising its validity. Statistical methodology for group sequential trials is well established when the endpoint is observed immediately, but less so for endpoints that are measured with a delay, such as repeatedly measured outcomes for which the primary measurement of interest is taken after several weeks. The latter can result in “pipeline” subjects at an interim analysis, these subjects may have early outcome measurements available, but their final endpoint is yet to be observed. Accounting for these early measurements has been shown to increase statistical power. Most importantly, pipeline patients will contribute with additional data after a decision to stop enrollment has been taken at an interim analysis. To make the best use of all available data, these data are ideally incorporated in the final analysis in a formal way. In this talk, we provide guidance on how to plan a GST with repeated measurements and a delayed endpoint and how to analyze the data resulting from these trials. We discuss three existing methods as proposed by Hampson and Jennison (2013, JRSS B: 75(1):3-54) and Jennison (2022, DSBS Course slides), as well as expand on them, for example adding a non-binding stopping rule for futility for the error spending methods proposed by Hampson and Jennison (2013, JRSS B: 75(1):3-54). The methods are illustrated using a case study. We discuss the pros and cons of each method and present results from a simulation study comparing method performance. The methods discussed have been implemented in R with code being freely available on Github.

Automation tools for group sequential designs with MTP and SSR

Yevgen Tymofyeyev, Michael Grayling
Johnson and Johnson, United States of America; ytymofye@its.jnj.com

Relatively recently methodology has been developed for strict error control in group sequential trials that test multiple endpoints. Amongst now available methods for this problem, the use of graphical testing procedures in group sequential designs offers a flexible and easy-to-implement design option. In this presentation, a case study will highlight the practical implementation and benefits of these approaches in designing rigorous and efficient clinical trials. As an implementation framework we describe how to combine the functionality of the popular R packages gMCP (for graphical testing) and gsDesign (for group sequential trials).
Finding a preferred group sequential design (including designs with SSR) is essentially a multi-parameter optimization problem, where a planner comes up with the optimal values for a set of control parameters, such as stage sizes (number of subjects or events for time-to-event setting), earlier stopping criteria and re-estimation rules. We demonstrate a process that can guide the selection of a preferred option that meets study-specific objectives and constraints. The process involves a 'try-and-error' search that is influenced by a formal optimization problem solving. To help expedite the utilization of the process in practice, we also illustrate newly developed code for the automation of this workflow as well.

Optimal drop-the-loser trials when an intermediate endpoint is used for interim selection

Samuel Sarkodie[1], James Wason[2], Michael Grayling[3]
[1]Newcastle University, United Kingdom; [2]Newcastle University, United Kingdom; [3]Janssen R&D, United Kingdom; s.k.sarkodie2@newcastle.ac.uk

The study proposes an integration of seamless phase II/III and drop-the-loser designs into one framework. Within this design, two treatments were compared to a shared control in phase II and the least effective arm was dropped from the subsequent phase III stage. To make good use of limited resources, an intermediate endpoint, envisioned to be cheaper and faster to use, was utilised at the first stage of the trial to inform adaptations before the definitive outcome was evaluated in the second stage.
The methodology accommodates the intermediate and definitive endpoints at the first and second stages respectively, while employing normal outcomes in both stages. Considering that adaptive designs rely on data available at the interim analysis to inform adaptations, an optimal timing for the interim analysis was proposed. Additionally, we show the scenario which maximises the family-wise error rate (FWER).
The key finding was that conducting the interim analysis when 65% of the data had been collected was optimal for cases where the correlation between the endpoints is treated as unknown in the FWER control requirement. However, in the case of treating as known for the FWER control requirement, a higher assumed value results in an earlier optimal timing for the interim analysis.

In conclusion, the proposed design accelerates the drug development process by integrating two traditionally separate trials and demonstrates efficiency in sample size. Compared to traditional non-ADs, the reduced sample size and the use of a less expensive endpoint to identify the treatment arm for subsequent data collection could lead to substantial cost savings.

Session VI: Designs with time-to-event endpoints

Friday, 26th April, 11:20 - 12:50

Sample size adaptations under non-proportional hazards using the restricted mean survival time

Carolin Herrmann[1], Paul Blanche[2]
[1]Charité - University Medicine Berlin, Institute of Biometry and Clinical Epidemiology, Germany; [2]University of Copenhagen, Section of Biostatistics, Denmark;
carolin.herrmann@posteo.de

In many clinical trials, the time until a specific event is of primary interest. When delayed treatment effects are prevalent and violate the proportional hazards assumption, the frequently applied hazard ratio is not straightforward to interpret. Hence, the restricted mean survival time has been increasingly promoted as an alternative effect measure in those scenarios. Another problem irrespective of the level of measurement is dealing with the insecurity in regards to the assumed effect size when it comes to sample size planning of a clinical trial. Hence, we propose a method for mid-trial sample size adaptations in adaptive group sequential designs when the restricted mean survival time is considered as primary endpoint. The sample size adaptation is supposed to happen at a fixed calendar time and is based on the conditional power. Special emphasis is placed on the pipeline data coming from the first stage, which contribute as left-truncated data [1] to a potential second stage. We consider three different approaches for including the left-truncated data into the inverse normal combination test (e.g. [2]). We use the inverse normal test to combine the data of the two stages. Performance of the approaches is evaluated and compared in terms of power and sample size. Moreover, we illustrate the method by a cardiologic example from biostatistical consultation. The new method extends the classic group sequential approach for the restricted mean survival time [3] and allows also for other adaptations than the sample size.

[1] N. Keiding, T. Bayer, S. Watt-Boolsen, Confirmatory analysis of survival data using left truncation of the life times of primary survivors, Stat Med, 1987, 6(8):939-944.
[2] Desseaux K, Porcher R. Flexible two-stage design with sample size reassessment for survival trials. Stat Med, 2007, 26(27):5002–5013.

[3] Y. Lu, L. Tian, Statistical Considerations for Sequential Analysis of the Restricted Mean Survival Time for Randomized Clinical Trials, Stat Biopharm Res, 2021, 13(2):210-218.

Blinded-into-unblinded interim analyses for time-to-event-trials with fixed analysis timings: a simulation study

Jan Meis[1], Carolin Herrmann[2], Björn Bokelmann[2], Meinhard Kieser[1]
[1]Institute of Medical Biometry, University of Heidelberg, Heidelberg, Germany; [2]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, Germany;
meis@imbi.uni-heidelberg.de

In time-to-event trials, the amount of information is determined by the number of events, not the number of recruited patients. For this reason, it is common that analysis timings are planned relative to reaching a threshold for the number of observed events. This makes the timing of the analyses random. In practice, there are many situations where it is much more convenient to plan analyses for a fixed calendar time rather than a random time point. For example, the trial might be sponsored by a federal funding agency that only funds trials for a certain number of years, or budgeting concerns of a pharmaceutical company might prohibit an excessively long trial. It might also be the case that current events render the trial results irrelevant if they are delayed for too long, which was a frequent concern during the COVID-19 pandemic. Further, coordination of data cleaning efforts for interim analyses in multi-center trials is much easier if the timing is known in advance. The main drawback of planning analyses for a fixed point in time is that the observed information level may be lower than expected, leading to a loss of power. This is especially concerning when estimates of the event rates are subject to a lot of uncertainty.
To deal with these issues, a pragmatic trial design for time-to-event trials with fixed analysis timings is proposed. This design features a blinded interim analysis at a fixed calendar time to estimate the event rate. If the observed number of events differs greatly from what was assumed, the data will be analyzed again in an unblinded interim analysis conducted by an independent statistician. This unblinded analysis assesses whether continuation of the trial as planned would be futile and whether this could be mitigated via design adjustments. In the other case, when the blinded analysis shows that the number of observed events is close to the planned number, the trial is continued without unblinding. This way, the potential for biasing investigators and the statistical dependency complications that come with unblinded interim looks are avoided.
We present a simulation study comparing various operational characteristics of this proposed trial design to more traditional approaches. We also discuss how the timing of the interim analysis can be determined and how appropriate thresholds for the blinding vs. unblinding decision can be derived.

Surviving the multiple testing problem: RMST-based tests in general factorial designs

Merle Munko[1], Marc Ditzhaus[1], Dennis Dobler[2], Jon Genuneit[3]
[1]Otto-von-Guericke University Magdeburg, Germany; [2]TU Dortmund University, Germany;
[3]Leipzig University, Germany; merle.munko@ovgu.de

Several methods in survival analysis are based on the proportional hazards assumption. However, this assumption is very restrictive and often not justifiable in practice. Therefore, effect estimands that do not rely on the proportional hazards assumption, such as the restricted mean survival time (RMST), are highly desirable in practical applications. The RMST is defined as the area under the survival curve up to a prespecified time point and, thus, summarizes the survival curve into a meaningful estimand. For two-sample comparisons based on the RMST, there is an inflation of the type-I error of the asymptotic test for small samples and, therefore, a two-sample permutation test has already been developed. The first goal is to further extend the permutation test for general factorial designs and general contrast hypotheses by considering a Wald-type test statistic and its asymptotic behavior. Additionally, a groupwise bootstrap approach is considered. In a second step, multiple tests for the RMST are developed to infer several null hypotheses simultaneously. Hereby, the asymptotically exact dependence structure between the local test statistics is incorporated to gain more power. The small sample performance of the proposed global and multiple testing procedures is analyzed in simulations and finally illustrated by analyzing a real data example.

Interim adaptations of weights in survival testing procedures

Moritz Fabian Danzer[1], Ina Dormuth[2]
[1]University of Münster, Germany; [2]TU Dortmund, Germany;
moritzfabian.danzer@ukmuenster.de

The use of weighted log-rank tests and combination tests based on them has been discussed intensively in the recent past, also due to corresponding use cases. There are far-reaching theoretical results on the optimality of certain procedures and extensive simulation studies in which different approaches are compared with each other. It is obvious that the selection of a test that is well suited or even optimal in the respective scenario can lead to power gains or savings in sample size.
However, at the time of planning a study, it is usually still unclear what form the effect will take and which test procedure would be optimal. It is therefore almost impossible to answer the question about the choice of method at this actually relevant point in time.
Adaptive designs now make it possible to inspect not only the strength of an effect but also the type of effect at the time of an interim analysis. As this is decisive for the choice of the test procedure, the situation should now be re-evaluated at this point. In this talk, we will look at ways to make meaningful use of the above-mentioned findings on the choice of weight. In particular, it will be investigated to what extent combination tests with a broader power function can be used in the first stage under uncertainty about the type of effect and how a decision can be made at the time of an interim analysis about the test with which the study should be continued.

CompAREdesign: An R Package for the Design of Randomized Clinical Trials with Composite Endpoints

Jordi Cortés Martínez[1], Marta Bofill Roig[2], Guadalupe Gómez Melis[1]
[1]Universitat Politècnica de Catalunya, Spain; [2]Medical University of Vienna, Austria;
jordi.cortes-martinez@upc.edu

Composite endpoints (CE) are defined as the occurrence of any of the relevant events in trials with binary response and as the time from randomization to the first observed event among all components in time-to-event studies. Using a CE as the primary endpoint could provide a broader overview of an intervention's efficacy and, if the event rates are too low, it could increase the study power.
We present the R package CompAREdesign designed to calculate the required sample size in randomized controlled trials with CE. Designing trials with time-to-event endpoints can be particularly challenging because the proportional hazard (PH) assumption usually does not hold when using a composite endpoint. Consequently, the conventional formulae for sample size calculation no longer apply. CompAREdesign calculates sample sizes in this situation based
on the log rank test and when the PH assumption holds for their components. The sample size is calculated on the basis of anticipated information on the composite components and the correlation between them.
CompAREdesign also includes functions to calculate the probability of observing the CE (for binary outcomes) and the survival function of the CE. In addition, the package can provide the expected treatment effect in terms of hazard ratio over time or geometric average hazard ratio for time-to-event studies, and as risk difference, relative risk or odds ratio for binary outcomes. To assess the statistical efficiency of employing CE as the primary endpoint, the Package provides users with the Asymptotic Relative Efficiency measure.
Beyond its quantitative functionalities, CompAREdesign incorporates a data simulation function, permitting researchers to simulate data with CE. This feature enhances the understanding of trial characteristics, such as power and type 1 error rate, according to the trial design assumptions and facilitates the refinement of study designs.
Although there are several R packages for the analysis of trials with CE, to our knowledge, the CompAREdesign package is the first specifically implemented to address the design of an RCT and as such CompAREdesign stands out as a valuable tool.

Session VII: Multiple regression

Friday, 26th April, 14:00 - 15:12

Assessing the role of interim analyses in addressing replication concerns in behavioral sciences

Beatrijs Moerkerke[1], Tom Loeys[1], Kelly Van Lancker[2]
[1]Department of Data-Analysis, Ghent University, Belgium; [2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium;
beatrijs.moerkerke@ugent.be

The replication crisis has led to widespread concerns across various scientific disciplines, leading to numerous evaluations of research practices. Within the field of psychology, replication initiatives have revealed that many findings are not replicable, confirming the necessity for heightened methodological rigor. In response, statisticians and methodologists are actively engaged in developing and refining statistical methods to ensure the accuracy and reliability of research findings.
In this pursuit, a recognized challenge in behavioral sciences is the presence of bias in effect estimation. Attention has shifted towards advocating for sufficiently powered studies, as underpowered studies can lead to an inflation of effect sizes in the scientific literature. Transitioning to more cost-effective flexible designs, such as adaptive and group sequential designs, presents a promising avenue to enhance the feasibility of attaining greater statistical power. It is well acknowledged that when a study with interim analyses stops early for efficacy, effect estimates may also become overestimated and should be corrected. It is crucial to understand the impact of this in fields with a lesser emphasis on replication. In these contexts, it is also important to consider the predictive value of study results. Therefore, we conduct a simulation to study these aspects in varying settings by creating scenarios where studies terminated early exert substantial influence in shaping the literature.

Sample Size Re-Estimation as an Alternative to A Priori Power Analysis in Social Sciences: The Multiple Linear Regression Case

Lara Vankelecom, Ole Schacht, Tom Loeys, Beatrijs Moerkerke
Department of Data-Analysis, Ghent University, Belgium; lara.vankelecom@ugent.be

It is widely acknowledged that research should be sufficiently powered for the results to be reliable. In psychological research, determining the sample size of the study by performing an a-priori power analysis is currently recognized as the best option to ensure sufficient statistical power. However, without any study data, accurately specifying the nuisance parameters necessary for the a-priori sample size calculation becomes a formidable challenge. Reliable estimates are seldom available beforehand since they are rarely reported in related published research. Using inaccurate values for the nuisance parameters

to calculate the required sample size has large consequences on the final power of the study. One way to avoid this issue is by switching from the traditional fixed design (where sample size needs to be determined beforehand) to an adaptive design (where sample size can still be modified through the course of the study). In this paper, we introduce one such adaptive design, known as the sample size re-estimation design. In this design, the first batch of collected data is used to estimate the nuisance parameters, which are subsequently employed to update the required sample size of the study. While this design is already thoroughly investigated for the independent-samples t-test, we explain and examine this design for multiple linear regression (where the nuisance parameters do not only include the variance, but also the correlation between the covariates). For this purpose, we construct a sample size formula for sample size calculation in this context. This formula estimates the sample size that is required to detect the effect of one predictor with pre-specified power, while controlling for other covariates. We investigate through simulation the implications of sample size re-estimation on the type I and type II error rate of the final test. Inflation of the type I error rate can be substantial when the interim estimates of the nuisance parameters are based on a small sample, however, approaches the nominal level when more “first batch” data are collected. Also, the power reaches on average the desired level when the interim estimates are based on a sufficiently large sample.

Information-Based Monitoring as an Alternative to A Priori Power Analysis in Social Sciences: The Multiple Linear Regression Case

Ole Schacht, Lara Vankelecom, Tom Loeys, Beatrijs Moerkerke
Department of Data-Analysis, Ghent University, Belgium; ole.schacht@ugent.be

Testing theory-based hypotheses lies at the core of social sciences. Within the frequentist framework, good scientific practice requires specifying a priori the power with which the presumed effect is to be detected if it truly exists. Unfortunately, deriving the sample size necessary for a given power is not always trivial. Indeed, multiple review papers revealed that power is frequently neglected in social sciences. Considering the replication crisis, underpowered studies have been identified as a key issue perpetuating the publication bias. Hence, making power analyses more feasible to conduct is of interest to social scientists. Since social scientists are faced with less stringent blinding procedures than in clinical trials and often have permanent access to the accumulating data, it will be demonstrated in this talk how researchers in social sciences in particular can benefit from using interim Fisher Information, as opposed to a priori power analysis. Fisher Information is defined as the reciprocal of the precision of the effect of interest. And so, the power problem can be reframed as specifying a sufficiently large Information criterion. Using asymptotics, it can be shown that this criterion depends only on the magnitude of the presumed effect, the significance level, and the power. The idea is then to monitor the Information during data accumulation and to terminate the data collection as soon as the Information exceeds the specified threshold. When this decision rule is adhered to, it can be shown that the desired power is attained, while the type-I error rate keeps its nominal level.
Focus will be on the practical usability of the method, and to a lesser extent on the technical details. The method will be illustrated in the context of multiple linear regression analysis. The use of Information-based monitoring brings potential as deriving the sample size in this

context may quickly become intractable. Indeed, when the focus is on a particular effect, but other nuisance parameters are also present, it is often not feasible to make an informed guess on the latter and to estimate a correct sample size prior to the start of the study. The Information-based approach may then be a viable alternative by, after having specified an initial guess on the sample size, monitoring the data until the desired power is reached. In this talk, an approach will be proposed to predict the final sample size during the study based on the observed Information fraction.

Multiple marginal models for multinomial regression with high-dimensional covariates

Thorsten Dickhaus, Vladimir Vutov
University of Bremen, Germany; dickhaus@uni-bremen.de

Modern high-throughput biomedical devices routinely produce data on a large scale, and the analysis of high-dimensional datasets has become commonplace in biomedical studies. However, given thousands or tens of thousands of measured variables in these datasets, extracting meaningful features poses a challenge. We propose a procedure to evaluate the strength of the associations between a nominal (categorical) response variable and multiple features (covariates) simultaneously. Specifically, we propose a framework of large-scale multiple testing under arbitrary correlation dependency among test statistics. First, marginal multinomial regressions are performed for each feature individually. Second, we use an approach of multiple marginal models for each baseline-category pair to establish asymptotic joint normality of the stacked vector of the marginal multinomial regression coefficients. Third, we estimate the (limiting) covariance matrix between the estimated coefficients from all marginal models. Finally, our approach approximates the realized false discovery proportion of a thresholding procedure for the marginal p-values for each baseline-category logit pair. We demonstrate a practical application of the method to hyperspectral imaging data. The dataset is obtained by a matrix-assisted laser desorption/ionization instrument. The presentation is based on Vutov and Dickhaus (2023a, https://doi.org/10.1002/bimj.202100328) and on Vutov and Dickhaus (2023b, https://doi.org/10.1002/sim.9761).

Invited Session II: Practical experiences of using software to design clinical trials using simulations

Friday, 26th April, 16:00 - 17:20

Using the "SIMulating PLatform trials Efficiently" (SIMPLE) R package to develop a simulator for a bespoke platform trial

Peter Jacko[1,2]
[1]Berry Consultants, United Kingdom; [2]Lancaster University, United Kingdom;
peter.jacko@gmail.com

EU-PEARL was a strategic partnership between the public and private sectors under the umbrella of the Innovative Medicines Initiative operating in 2019-2023 to shape the future of clinical trials by developing a generic framework and a set of tools to conduct patient-centric collaborative platform trials. One of the working packages of the EU-PEARL project was dedicated to a rare disease called neurofibromatosis type 2 (NF2) and associated progressive tumors. We set out to develop a design of a platform trial which could be replicable for similar rare diseases, which required a simulator to evaluate the operating characteristics and thus properly understand the influence of design features on the performance of the design. For that end we used the "SIMulating PLatform trials Efficiently" (SIMPLE) R package, which is a modular tool for simulating complex platform trials developed within the scope of EU-PEARL. In this talk we will present our experience with developing an extension of the SIMPLE simulator adapted to the particular setting of a rare disease. We will focus on lessons learned which are transferable to development of simulators for other bespoke platform trials.

Flexible Clinical Trial Planning with the R Package rpact

Gernot Wassmer, Friedrich Pahlke
RPACT, Germany; friedrich.pahlke@rpact.com

In our upcoming presentation, we will explore the capabilities of the validated open-source R package 'rpact,' which is available on CRAN and GitHub. This package is useful for conducting flexible simulations in clinical trial planning. As an example, we will demonstrate how 'rpact' enables users to easily define new functions for calculating the number of subjects or events required, based on given conditional power and critical values for specific testing scenarios. This includes the implementation of advanced strategies like the 'promising zone approach'.
Drawing from our own experiences, we will also discuss how we successfully met a stringent FDA deadline by employing parallelized simulations with 'rpact', ensuring both efficiency and reliability in a 'seed' safe manner.

Furthermore, we introduce 'RPACT Cloud', a user-friendly platform designed to simplify and enhance the process of clinical trial simulations for researchers and practitioners.
This session is an opportunity to discover how 'rpact' and 'RPACT Cloud' can streamline your clinical trial planning and execution.

Design or simulation: What comes first?

Tobias Mielke
Johnson & Johnson, Germany; tmielke1@its.jnj.com

Adaptive trial designs are utilized in clinical research to increase efficiency and to mitigate uncertainties which may exist in the planning stage of clinical trials. The adaptive interim decision rules limit the applicability of standard sample size formulae for such trials. Simulations are used in the planning stage to assess and optimize operating characteristics of such adaptive clinical trials. Once simulations are completed, a range of operating characteristics are typically summarized, such as expected sample size, type-1 error, power, time-to-decision, probability of false futility decisions. All those operating characteristics are assessed under a range of scenarios on the true underlying distribution of efficacy parameters. Having such a multitude of operating characteristics results in a multiplicity problem in design optimization: designs may win on one characteristic for one scenario, while losing on others, such that design discussions could be easily mislead by “false-positives”, or might become highly subjective. Not surprisingly, an appropriate “simulation study process” would typically resemble conventional research processes, which would also be applied within clinical drug development.
The simulation design starts with a proper problem definition, reflecting the simulation study objective. What follows is a definition of the relevant scenarios and the candidate designs. Those two resemble the “populations” and “study groups” within clinical trials. Finally, you would like to compare the operating characteristics of study designs under different scenarios. The operating characteristics as such take the role of “endpoints”. As in drug development, you could eventually start with an unconstrained “design screening process”, prior to moving into the “confirmatory” phase of design selection. To achieve a successful outcome of this design screening process, a “target-design-profile” should optimally be available. Otherwise, design optimization may result in overly complex solutions with limited added utility (vs. simpler ones). In all efforts, one should ensure that the simulation studies follow “GCP”-principles to provide valid results. Validated off-the-shelf simulation software allows to do so in a rapid, reproducible manner. Still, sufficient time should be invested to understand and communicate the simulation outputs. Even the best adaptive design may not be implemented, if it merits can’t be clearly explained to the decision makers, who are frequently no statisticians.
In this presentation, I will share experiences on the processes of designing adaptive clinical trials, as well as the process of designing and testing simulation software.