

Analysis of Two-Stage Seamless Adaptive Design in Clinical Development and Its Applications

by

Weijia Mai

Department of Biostatistics & Bioinformatics
Duke University

Defense Date: October 28, 2025

Approved:

Shein-Chung Chow, Advisor

Sarah Peskoe

Frank W. Rockhold

Ayako Suzuki

Yuan Wu

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy
in the Department of Biostatistics & Bioinformatics in The Graduate School of
Duke University
2025

ABSTRACT

Analysis of Two-Stage Seamless Adaptive Design in Clinical Development and Its Applications

by

Weijia Mai

Department of Biostatistics & Bioinformatics
Duke University

Defense Date: October 28, 2025

Approved:

Shein-Chung Chow, Advisor

Sarah Peskoe

Frank W. Rockhold

Ayako Suzuki

Yuan Wu

An abstract of a thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics & Bioinformatics in The Graduate School of Duke University
2025

Copyright by
Weijia Mai
2025

Abstract

In recent years, innovative two stage seamless adaptive designs are increasingly used in clinical development to accelerate decision making and improve efficiency while preserving confirmatory integrity. This thesis develops methods for two stage seamless adaptive trials in which interim and final stages may use different endpoints, and the enrolled population may shift over time. It introduces a design level mean framework that links early surrogate markers readouts to later primary clinical endpoints under a common objective, and derives a surrogate informed combined estimator for paired, partially overlapping and independent settings. Under certain conditions, the estimator is unbiased, uses variance minimizing weights, and applies correlation aware spending to preserve overall type I error.

The second part of this thesis addresses efficient evidence for biosimilarity assessment. An innovative Biosimilarity Index (BI) is introduced to formalize decision making for biosimilarity when pharmacokinetic similarity provides the primary evidentiary basis, with performance compared against the conventional confidence interval approach. The idea of BI was then extended to an innovative Relative Biosimilarity Index (RBI) that explicitly incorporates inherent variability and provides a method to define a proposed regulatory similarity threshold for RBI testing that is interpretable across products.

Collectively, these contributions expand novel statistical methods for two-stage seamless adaptive designs and biosimilarity evaluation, delivering procedures that are statistically accurate, efficient and aligned with practical development and review needs.

Contents

Abstract.....	iv
List of Tables	x
List of Figures.....	xii
Acknowledgements.....	xiii
1. Introduction.....	1
1.1 Background.....	1
1.1.1 Complex Innovative Design.....	1
1.1.2 Two-Stage Seamless Adaptive Design	12
1.2 Types of Two-stage Seamless Adaptive Designs and Current Analysis Methods	14
1.2.1 Types of Two-Stage Seamless Adaptive Designs.....	15
1.2.2 Statistical Challenges and Literature Review of Two-Stage Seamless Adaptive Designs	16
1.2.2.1 Different Study Objectives	18
1.2.2.2 Different Study Endpoints.....	19
1.2.2.3 Different Study Populations	20
1.3 Research Gap and Problem Statement.....	22
1.3.1 Shifts in Study Population and Change in Study Endpoints	22
1.3.2 Challenges in Biosimilar Development.....	23
1.3.3 Lack of Similarity Threshold Frameworks for Biosimilar Development	24
1.4 Research Objectives	25
1.4.1 Development of a Statistical Method for Analysing Two-Stage Seamless Adaptive Designs with Population Shifts and Endpoint Changes.....	26
1.4.2 Proposal for the Use of a Two-Stage Seamless Adaptive Design in Biosimilar Product Development	26

1.4.3	Extension: Development of a Similarity Threshold Framework Using the Relative Biosimilarity Index in Biosimilar Product Development	27
1.5	Significance of the Research	28
1.6	Thesis Organization.....	30
2.	Analysis of Innovative Two-stage Seamless Adaptive Design with Different Endpoints and Population Shift	32
2.1	Introduction	32
2.2	Background and Motivation	35
2.3	Method of Data Combination when both Shift in Population and Endpoint Difference Occur	36
2.3.1	Step 1: Interim Analysis at the End of Stage 1 (before protocol amendment).....	36
2.3.2	Step 2: Collect Information at the End of Stage 2.....	38
2.3.3	Step 3: Consider Population Shift (mean-level model, per arm)	39
2.3.4	Step 4: Hypothesis Testing at Stage 2 (post-shift treatment effect)	40
2.4	Type I Error Control.....	41
2.4.1	Setup and Notation.....	41
2.4.2	Estimating ρ in Practice	43
2.4.2.1	Design-Based plug-in (when Stage 2 reuses Stage 1).....	43
2.4.2.2	Pilot-based Monte Carlo.....	43
2.5	Sample Size Calculation.....	44
2.6	Simulation Study and Sensitivity Analysis.....	45
2.7	Conclusion and Discussion.....	51
3.	A Proposal for the Use of Innovative Two-Stage PK-Clinical Adaptive Trial in Biosimilar Product Development	53
3.1	Introduction	54
3.2	Biosimilarity Index.....	57

3.2.1	Derivation of Biosimilarity Index and Statistical Properties.....	58
3.2.2	Remarks	63
3.3	Proposed Two-stage PK-Comparative Clinical Adaptive Trial	64
3.3.1	Analysis for Stage 1	65
3.3.2	Analysis for Stage 2	65
3.3.2.1	Endpoints Relationship is Not Well-established	66
3.3.2.2	Endpoints Relationship is Well-established	67
3.4	Type I Error Control	70
3.4.1	Setup and Notation.....	70
3.4.2	Estimating ρ in Practice	71
3.4.2.1	Design-Based plug-in (when Stage 2 reuses Stage 1)	71
3.4.2.2	Pilot-based Monte Carlo.....	72
3.5	Simulation.....	72
3.5.1	Corresponding p_0^* at Different Sample Sizes when Using CI Approach	73
3.5.2	Comparing Conclusions on Biosimilarity with Different Approaches	74
3.5.3	Varying Sample Sizes	75
3.5.4	Determining the Appropriate Sample Size for BI Approach Given p_0	77
3.5.5	Determining the Appropriate p_0 for BI Approach Given Sample Size.....	80
3.6	Discussion and Concluding Remarks	84
4.	An Innovative Method Based on Relative Biosimilarity Index for Assessing Biosimilar Drug Products	88
4.1	Introduction	89
4.2	Relative Biosimilarity Index (RBI)	92
4.2.1	Biosimilarity Index.....	92
4.2.2	Estimation of P_{RR}	93

4.3	Statistical Testing Methods for Relative Biosimilarity Index (RBI): A Comparative Framework	95
4.3.1	Fieller's Test.....	95
4.3.1.1	Test Statistic	96
4.3.1.2	Variance and Covariance Estimation	97
4.3.1.2.1	Delta Method Estimation.....	97
4.3.1.2.2	Bootstrap-Based Estimation	97
4.3.2	Log Transformation	98
4.3.3	Regression Inference with Bootstrap Sampling	99
4.3.3.1	Testing Mean RBI	100
4.3.3.2	Testing $Pr(RBI > \delta) \geq 1 - \alpha$	101
4.3.4	Comparison of Testing Methods.....	103
4.4	Method for Selecting δ in Testing the Relative Biosimilarity Index (RBI)	106
4.4.1	Method	107
4.4.2	Choosing the Tuning Parameter γ	109
4.5	Simulation Study	110
4.5.1	Determination of γ	113
4.5.2	Simulation-Based Evaluation of Sample Size and Type I Error Control.....	118
4.5.2.1	Power-Based Sample Size Guidance.....	118
4.5.2.2	Type I Error Evaluation and Sample Size Validation	119
4.5.3	Method-Specific Performance Evaluation under Selected Threshold.....	121
4.6	Discussion and Concluding Remarks	125
5.	Conclusion and Future Directions	128
5.1	Integration of Contributions	129

5.2 Innovation and Impact	131
5.3 Limitations.....	133
5.4 Future Directions	134
5.5 Concluding Remarks	135
Appendix A – Unbiasedness and Minimum Variance of Combined Estimator	137
Appendix B – Overlap-Specific Variance Derivations (per arm).....	139
Appendix C – Unbiasedness of the PK-Informed Estimator	142
References	143

List of Tables

Table 1: Common types of adaptive designs.....	7
Table 2: Types of two-stage seamless adaptive designs (depending upon objective, endpoint, and population)	15
Table 3: Statistical tests for various hypotheses testing at stage 1.....	38
Table 4: Decision boundaries at the end of stage 1.....	38
Table 5: Information collected at each stage.....	39
Table 6: Statistical tests for various hypotheses testing at stage 2.....	40
Table 7: Decision boundaries at the end of stage 2.....	41
Table 8: Sample size requirement for various hypotheses testing.	44
Table 9: Power with different allocation ratio when sample size per arm is 100.....	47
Table 10: Power with different allocation ratio when sample size per arm is 50.....	47
Table 11: Power with different allocation ratio when sample size per arm is 30.....	47
Table 12: Power with different allocation ratio when sample size per arm is 30.....	48
Table 13: Power with different allocation ratio when sample size per arm is 100, effect size 1.5.	49
Table 14: Power with different allocation ratio when sample size per arm is 100, effect size 1.4.	50
Table 15: Characteristics of two-stage PK-Clinical proposal.	64
Table 16: Decision boundaries at the end of stage 1.....	65
Table 17: Decision boundaries at the end of stage 2.....	67
Table 18: Corresponding p_0^* of <i>CI</i> approach.....	73
Table 19: Biosimilarity concluded from <i>CI</i> and <i>BI</i> approach with varying p_0	74
Table 20: Simulation results with varying sample sizes.	75
Table 21: Probability of success of the trial with varying p_0 and sample sizes.	81

Table 22: Comparative summary of six statistical methods..... 103

Table 23: Empirical minimum sample size for achieving sufficient power (80%) for varying δ , with fixed coefficient of variation ($CV = 0.20$) and mean difference ($meandiff = 0.1$)..... 119

Table 24: Empirical type I error rates under the null hypothesis for varying sample sizes, with fixed coefficient of variation ($CV = 0.20$) and mean difference (0.4). 120

List of Figures

Figure 1: Illustration of two-stage seamless design.....	36
Figure 2: Simulation results with sample size per arm 100.....	46
Figure 3: Simulation results with sample size per arm 50.....	46
Figure 4: Simulation results with sample size per arm 30.....	47
Figure 5: Simulation results with sample size per arm 30.....	48
Figure 6: Power change when sample size per arm 100, effect size 1.5.....	49
Figure 7: Power change when sample size per arm 100, effect size 1.4.....	50
Figure 8: Simulation results when $p_0 = 0.4$	77
Figure 9: Simulation results when $p_0 = 0.5$	78
Figure 10: Simulation results when $p_0 = 0.6$	78
Figure 11: Simulation results when $p_0 = 0.7$	79
Figure 12: Simulation results when $p_0 = 0.8$	79
Figure 13: Power vs. CV for different γ values.....	114
Figure 14: Power vs. γ for different CV levels.....	115
Figure 15: Type I error vs. CV for different γ values.....	116
Figure 16: Type I error vs. γ for different CV levels.....	117
Figure 17: Power with different sample sizes and δ at $CV = 0.2$	119
Figure 18: Type I error rate as a function of sample size per group for fixed coefficient of variation ($CV = 0.20$) and decision threshold ($\delta = 0.8$).....	120
Figure 19: Type I error and power across varying sample size for Fieller's method, beta regression and logit transformation tests.....	123
Figure 20: Type I error and power across varying sample size for probability-based tests.....	123

Acknowledgements

I would like to extend my deepest gratitude to my advisor, Professor Shein-Chung Chow, for his steadfast guidance, rigorous standards, and generous mentorship throughout my doctoral training. His commitment to clarity, practicality, and scholarship has profoundly shaped this dissertation and my development as a researcher and biostatistician.

I am equally grateful to the members of my dissertation committee for their thoughtful feedback and constructive suggestions at every stage. Their careful reading and probing questions strengthened the quality and impact of this work.

My appreciation also goes to the faculty and staff of the Department of Biostatistics & Bioinformatics, whose teaching, administrative assistance, and encouragement have been invaluable. I am grateful to my peers for their support and for the many stimulating conversations that sharpened my ideas and sustained my motivation.

Finally, I owe heartfelt thanks to my family and friends for their patience, understanding, and unwavering support.

1. Introduction

1.1 Background

The increasing complexity of drug development and the demand for greater efficiency have underscored the importance of innovative trial designs. From a clinical perspective, there is a growing need to detect signals or trends of safety and efficacy more efficiently, enabling timely decisions about the future of a test treatment. From a regulatory perspective, novel designs must still generate substantial evidence to support the approval of new therapies. Regulatory agencies, such as the United States (US) Food and Drug Administration (FDA) and European Union (EU) European Medicine Agency (EMA), have issued multiple guidance documents to support the adoption of innovative approaches while maintaining compliance with regulatory standards. From a statistical perspective, it is essential that these designs ensure integrity, quality and scientific validity of clinical data, which are critical to both regulatory review and clinical interpretation. In response to these needs, seamless adaptive trial designs have emerged as a flexible and efficient alternative to traditional sequential designs, offering the potential to streamline clinical development while maintaining rigor and credibility.

1.1.1 Complex Innovative Design

In 2018, the US FDA recommended the use of complex innovative design (CID) for clinical investigation of a test treatment under study for a flexible, efficient, accurate and reliable assessment of the test treatment. The CID includes adaptive design, n-of-1 trial design, master protocol design, and Bayesian group sequential design. In recent years, the pursuit of more flexible, efficient, ethical, and informative clinical trials has led

to the emergence and growing adoption of flexible (adaptive) designs. These designs provide a flexible and rigorous framework for modifying key elements of a clinical trial based on accumulating interim data, all within a prospectively planned structure. Such flexibility is especially important in modern clinical development, where traditional fixed designs may not accommodate evolving knowledge or uncertainties encountered during trial conduct (see, e.g., Chow & Chang, 2011; FDA, 2019).

Adaptive trial designs are defined by the US FDA as “a clinical trial design that allows for prospectively planned modifications to one or more aspects of the design based on accumulating data from subjects in the trial” (FDA, 2019). Unlike post hoc protocol amendments, which are reactive and often require regulatory resubmission, adaptations are pre-specified in the protocol and guided by predefined decision rules. This approach ensures statistical validity while allowing the trial to respond dynamically to emerging information (FDA, 2019).

Key features that distinguish adaptive designs from traditional fixed designs include their ability to modify parameters such as sample size, treatment arms, randomization ratios, inclusion/exclusion criteria, and study endpoints. These adaptations are made based on interim analyses and are governed by statistical rules aimed at maintaining control of Type I error rate and achieving desired statistical power (or minimizing Type II error rate (Jennison & Turnbull, 2000; Chow & Chang, 2011). For example, a treatment arm demonstrating clear futility may be dropped early, or a trial may be expanded if early signs of efficacy are observed (FDA, 2019). Crucially, these

adaptations are implemented in a manner that preserves the integrity and interpretability of the final analysis.

Adaptive designs are particularly advantageous in settings where uncertainties exist at trial initiation, such as in early-phase trials, precision medicine applications, or rare disease studies. They offer several potential benefits over traditional designs:

- **Flexibility**

The defining advantage of adaptive designs is flexibility: prospectively specified modifications to aspects such as sample size, randomisation ratios, treatment arms/doses, or target subpopulations in response to interim information (Jennison & Turnbull, 2000; Bretz et al., 2009; Chow & Chang, 2011; Kairalla et al., 2012; Van der Baan et al., 2012; Wang et al., 2009; FDA 2019). This designed-in ability to learn and adapt during the trial is precisely what enables the ethical and economic efficiency gains described below.

- **Efficiency**

Adaptive designs may offer significant gains in operational and resource efficiency by enabling prospectively planned modifications to the trial based on accumulating interim data (Jennison & Turnbull, 2000; Kairalla et al., 2012; Van der Baan et al., 2012; Wang et al., 2009). One of the primary advantages lies in the ability to reduce development timelines and trial costs through early stopping for futility or efficacy. For instance, if interim results indicate that a treatment is unlikely to achieve its intended clinical benefit, the trial may be terminated early, thereby avoiding the enrolment of additional participants and conserving

resources. This may also minimize patient exposure to ineffective or potentially harmful therapies. Conversely, when interim analyses reveal strong evidence of treatment efficacy, the trial may be adapted to accelerate progression, expand enrolment, or reallocate resources toward further investigation. Such responsiveness enhances the overall efficiency of the development process, both ethically and economically.

- **Resource Optimization**

One of the core benefits of adaptive clinical trial designs is their ability to potentially enhance resource optimization. By incorporating procedures such as sample size re-estimation, interim analyses, and adaptive allocation strategies, adaptive designs allow investigators to dynamically adjust resource use according to accumulating evidence (Chow & Chang, 2011; Mehta & Pocock, 2011; Hatfield et al., 2016). For example, when early data indicate that a treatment arm is unlikely to demonstrate sufficient efficacy, that arm may be dropped, thereby avoiding further investment in unpromising options. Similarly, if variability estimates from interim data differ substantially from initial assumptions, sample size can be re-estimated to ensure adequate power without over-recruitment. Moreover, the sample size required may decrease when the adaptive designs such as seamless phase 2/3 design are applied appropriately (see e.g., Teng et al., 2020; Jiang & Yuan, 2023). These features contribute to a more efficient allocation of participants, funding, and time across the development program.

- **Ethical Advantage**

In addition, adaptive designs offer significant ethical advantages. The ability to terminate a trial early due to lack of efficacy, excessive toxicity, or overwhelming efficacy minimizes the exposure of participants to ineffective or harmful interventions (Pallmann et al., 2018; Gershon et al., 2023). This feature is particularly critical in areas such as oncology or rare diseases, where treatment options may be limited, and patient populations are vulnerable. Early stopping for harm prevents the continued administration of detrimental treatments, while early stopping for success allows beneficial therapies to reach broader patient populations more rapidly. Furthermore, adaptive designs may incorporate enrichment strategies to focus recruitment on patient subgroups showing greater likelihood of benefit, reducing unnecessary enrolment of patients unlikely to respond (Thall, 2021; Simon & Simon, 2013).

- **Improved Decision-making**

Adaptive designs can enhance the decision-making process in clinical trials by enabling interim evaluations of accumulating data, which can inform timely and evidence-based adaptations to the trial structure. Unlike traditional designs, which typically wait until the study's completion to assess efficacy or futility, adaptive trials incorporate pre-specified interim analyses that allow sponsors and investigators to make decision on adjustments. These may include increasing sample size, modifying randomization ratios, dropping or adding treatment arms, or stopping the trial early for efficacy or futility (Kairalla et al.,

2012; Pallmann et al., 2018; Mehta & Pocock, 2011). This flexibility supports timely decision-making and improves the efficiency of the trial, thereby accelerating the overall clinical development process.

Multi-arm adaptive platforms further support rapid go/no-go decisions by allowing simultaneous comparison of multiple treatments and adapting based on interim performance (Wason et al., 2012). In oncology and precision medicine applications, Bayesian adaptive designs enable continuous decision-making based on the strength of accumulating evidence, allowing for accelerated identification of promising therapies (Berry et al., 2013; Yada, 2022; Zhou et al., 2008). These features contribute to faster, more efficient clinical development by shortening the time needed to reach definitive conclusions or pivoting away from unpromising interventions.

- **Support for Precision/Personalized Medicine**

Adaptive designs have become an essential methodological tool in advancing precision medicine, which seeks to tailor therapies to clinically meaningful subgroups of patients defined by shared genetics, biomarkers, or phenotypic features. Among these designs, adaptive enrichment approaches are particularly well-suited to identify and focus on subpopulations most likely to benefit from treatment during the course of a trial (Thall, 2021). By using interim analyses to assess heterogeneity of treatment effects across biomarker-defined subgroups, adaptive trials enable the dynamic refinement of study populations in real time.

This capability is also relevant in oncology, where treatment response often varies based on molecular or genomic signatures. For example, Zhou et al. (2008) proposed a Bayesian adaptive design for the development of targeted therapies in non-small cell lung cancer, demonstrating how biomarker-guided adaptation can increase trial efficiency and align with the goals of personalized medicine. Their approach enables early identification of effective agents, while allowing the trial to eliminate the ineffective agents and therefore match effective treatments with patients' biomarker profile. Such designs reduce unnecessary exposure to ineffective interventions and accelerate the identification of promising effective treatment for each patient.

Recent methodological advancements have further expanded the scope of adaptive design, including in high-dimensional biomarker settings (Yada, 2022). These innovations enhance the ability of trials to adjust to emerging evidence, thereby aligning more closely with the goals of personalized healthcare. As healthcare continues to shift toward personalized treatment paradigms, adaptive designs, particularly those incorporating subgroup selection and biomarker-guided adaptations, are increasingly recognized as a foundational element of precision clinical research.

There are several categories of adaptive designs, each suited for different objectives. These include but are not limited to (Chow & Chang, 2008):

Table 1: Common types of adaptive designs

Design Type	Description	Example References
-------------	-------------	--------------------

Group Sequential Designs	Allows early stopping for efficacy or futility based on interim analyses.	Posch & Bauer (1999); Jennison & Turnbull (2000); Müller & Schäfer (2001); Jennison & Turnbull (2006)
Sample Size Re-estimation Designs	Adjusts sample size based on interim estimates of variance or effect size to maintain power.	Lehmacher & Wassmer (1999); Gao et al. (2008); Pritchett et al. (2015)
Response-Adaptive Randomization	Modifies allocation probabilities during the trial to favour better-performing treatments.	Rosenberger et al. (2001); Robertson et al. (2023)
Drop-the-Loser Designs	Eliminates inferior treatment arms as data accumulate.	Sampson & Sill (2005); Sun et al. (2006)
Seamless Phase II/III Designs	Combines exploratory (Phase II) and confirmatory (Phase III) phases into a single continuous trial.	Kelly et al. (2005); Chow et al. (2007)
Biomarker-Adaptive Designs	Allows adaptation based on biomarker data, including enrichment or subgroup targeting.	Wang et al. (2007); Yada (2022); Jin & Zhang (2023)

Despite their advantages, adaptive designs are associated with specific challenges. Their planning and implementation require sophisticated simulations to assess performance across various scenarios. The increased complexity in design

necessitates detailed documentation of adaptation rules, and real-time data monitoring must be carefully controlled to avoid operational bias. Statistically, maintaining Type I and Type II error control and avoiding inflated false-positive conclusions remain central concerns. Furthermore, while regulatory acceptance of adaptive designs has improved (FDA, 2019), some forms of adaptation still lack well-established methodological frameworks, limiting their broader application.

In summary, adaptive clinical trial designs offer a promising alternative to traditional fixed designs by incorporating flexibility and efficiency into trial conduct. When properly planned and executed, they have the potential to accelerate development timelines, optimize the use of limited resources, and improve outcomes for both sponsors and patients without compromising scientific validity. As regulatory agencies continue to provide guidance and statistical methodologies mature, adaptive designs are expected to play an increasingly vital role in the future of clinical research.

While adaptive designs offer substantial advantages in flexibility, efficiency, and ethical responsiveness, they also present several methodological and operational challenges that must be carefully addressed to preserve scientific validity and regulatory acceptability. The key limitations include design and implementation complexity, incomplete statistical support, and potential for operational bias.

- **Complexity in Design and Implementation**

Adaptive designs require more intricate planning and operational execution compared to traditional fixed designs. Each potential adaptation, whether involving changes in sample size, dose level, treatment arms, or

endpoints, must be rigorously pre-specified and statistically justified. The necessity to define clear and objective adaptation rules adds an additional layer of complexity to both the protocol and statistical analysis plan (Chow & Chang, 2011; Pallmann et al., 2018). Furthermore, interim analyses and real-time decision-making require specialized infrastructure, including secure data monitoring, access control procedures, and advanced software tools to simulate and validate adaptation procedures. These requirements often result in longer setup times, increased operational costs, and a greater demand for specialized statistical and operational expertise (Kairalla et al., 2012; Quinlan & Krams, 2006; Gallo et al., 2006).

- **Inadequate Statistical Support**

Despite the progress in adaptive methodology, not all adaptive designs are fully supported by established statistical frameworks. Unlike fixed designs, where inference procedures are straightforward and well-validated, adaptive trials introduce complexities related to error rate control, bias, and variance estimation. Mid-course adaptations, such as adding treatment arms or re-estimating sample size, can inadvertently inflate the Type I or Type II error rates if not appropriately adjusted for (Mehta & Pocock, 2011; Wassmer & Brannath, 2016). Moreover, statistical tools for certain complex adaptations, such as change in study endpoints or biomarker-driven dynamic enrichment, are still evolving and lack universally accepted implementation standards. This lack of methodological maturity may complicate interpretation, limit generalizability,

and pose potential challenges during regulatory review (FDA, 2019; Gallo et al., 2006).

- **Potential for Operational Bias**

The inherent flexibility of adaptive trials introduces increased susceptibility to operational bias, especially when interim analyses are not adequately blinded or when adaptations rely on subjective interpretation of early results (Pallmann et al., 2018). For instance, prematurely stopping a trial based on overly optimistic interim outcomes may result in overestimated treatment effects that do not replicate in subsequent studies. Moreover, trials with early stopping rules may fail to detect meaningful long-term outcomes (Gershon et al., 2023). Moreover, adaptive decisions, such as modifying randomization ratios or eligibility criteria, may be inadvertently influenced by factors such as operational considerations, commercial pressures, or investigator expectations, rather than pre-specified statistical criteria (Huskins et al., 2018; Pallmann et al., 2018). Such non-scientific influences can threaten the internal validity of the trial and undermine confidence in its conclusions. To mitigate these risks, adaptive trials must incorporate robust oversight mechanisms, including independent data monitoring committees, and strictly adhere to pre-planned decision rules and blinding protocols.

In summary, while adaptive designs provide a powerful and flexible framework for improving clinical trial efficiency, their successful implementation depends on meticulous planning, rigorous statistical methodology, and robust trial monitoring.

When limitations such as design complexity, incomplete statistical support, and operational bias are proactively addressed, adaptive trials can maintain scientific integrity and significantly enhance the pace and precision of drug development.

1.1.2 Two-Stage Seamless Adaptive Design

A two-stage seamless adaptive design represents a specific subclass of adaptive clinical trial methodologies, wherein two traditionally distinct trial phases, such as Phase 2 and Phase 3, are integrated into a single study. As described by Chow and Lin (2015), this approach refers to a design that combines two independent trials into one unified framework, thereby enabling the evaluation of objectives from both stages within a continuous development process. Data collected both prior to and following adaptation are incorporated into the final analysis, ensuring continuity in statistical inference and eliminating the need for initiating a separate follow-up trial between phases.

The principal objective of a two-stage seamless design is to streamline clinical development by removing the conventional pause that typically occurs between sequential phases. This facilitates a more efficient transition from early- to late-stage evaluation and allows for adaptive decision-making based on interim data collected during the first stage (Chow & Chang, 2011; Chow & Lin, 2015). At the interim analysis, typically conducted at the conclusion of Stage 1, accumulated data are evaluated to inform decisions regarding continuation, modification, or early termination of the trial. Crucially, the integrity of the trial is preserved by allowing data from both stages to be analysed jointly in the final analysis, thereby avoiding redundant regulatory

submissions and the logistical delays associated with initiating a new study (Chow & Chang, 2011).

This framework accommodates a range of pre-planned adaptations, including but not limited to treatment arm selection, sample size re-estimation, modification of eligibility criteria, adjustment of study endpoints, and early stopping for futility or efficacy. Such flexibility enables real-time trial optimization and may reduce development timelines and associated resource expenditure. For instance, treatments exhibiting promising efficacy signals during the interim analysis may seamlessly transition into confirmatory testing, whereas treatments demonstrating limited benefit may be discontinued. This adaptive capability facilitates efficient and ethical resource allocation (Pallmann et al., 2018; Sverdlov & Wong, 2014; Chow & Lin, 2015).

Given its potential to support early termination decisions based on strong evidence or futility, the two-stage seamless design can also expedite the transition to regulatory review, thereby accelerating the timeline from development to potential approval. These advantages are particularly relevant in **time-sensitive therapeutic areas**, such as infectious disease outbreaks (e.g., pandemics), where timely decision-making and efficient resource allocation are critical. Similarly, in **oncology**, this design has been used to accelerate the evaluation of treatments by using intermediate outcomes and/or simultaneously testing various new agents (Parmar et al., 2008).

Despite its operational and ethical advantages, the implementation of a two-stage seamless design necessitates rigorous pre-trial planning and advanced statistical methodology (Zhu & Wong, 2023). Since adaptations are informed by interim data,

failure to account for potential biases and proper control of the type I error rate may compromise the validity of the final analysis. To address these challenges, robust simulation studies and prospectively defined adaptation rules would be critical components of trial planning and execution (Pallmann et al., 2018).

In summary, the two-stage seamless adaptive design offers a strategically valuable framework for modern clinical research. By combining trial phases into a single protocol and incorporating adaptive elements that preserve statistical integrity, this design enhances development efficiency, promotes ethical and prudent use of resources, and aligns with the increasing demand for flexible, data-responsive methodologies in complex therapeutic areas.

1.2 Types of Two-stage Seamless Adaptive Designs and Current Analysis Methods

Two-stage seamless adaptive designs represent an advanced form of adaptive clinical trial methodology, enabling prespecified modifications between two stages of a trial, typically following an interim analysis of the data accumulated in the first stage. These designs allow for the integration of two clinical trial phases into a single, continuous study protocol, thereby improving efficiency of the trial, reducing costs, and enhancing ethical oversight. These designs can accommodate changes in key trial parameters, such as sample size, treatment dose levels, and patient populations, depending on the interim results. The ability to combine two trials that were planned and conducted independently, and to implement prespecified modifications, enhances

the trial’s adaptability and improves the efficiency while preserving scientific rigor and statistical validity.

1.2.1 Types of Two-Stage Seamless Adaptive Designs

The categorization of two-stage seamless adaptive designs primarily depends on the changes that occur between the two stages, specifically whether these involve the study's objectives, endpoints, or population. According to Chow (2020), these designs can be classified into eight categories based on whether the study objectives, endpoints, and population are the same or different at each stage. The designs are grouped as follows:

- i. 0-D design: No differences between stages (i.e., the same objectives, endpoints, and population).
- ii. 1-D design: One element (objectives, endpoints, or population) differs between stages.
- iii. 2-D design: Two elements differ between stages.
- iv. 3-D design: All three elements differ between stages.

Table 2: Types of two-stage seamless adaptive designs (depending upon objective, endpoint, and population)

	Target Patient Population	
	Same (S)	Different (D)
	Study Endpoints	Study Endpoints

Study Objectives	Same (S)	Different (D)	Same (S)	Different (D)
Same (S)	SSS	SDS	SSD	SDD
Different (D)	DSS	DDS	DSD	DDD

The complexity of the design increases as the number of differences ("D")

increases. As k increases from 0 to 3, the complexity of the study design grows, requiring more sophisticated statistical methodology to ensure proper control of Type I error and trial integrity across adaptations. Further classification based on endpoint data types can also be made under the scenario that study endpoints of two stages are different.

These different categories of two-stage designs are crucial for aligning trial conduct with evolving clinical needs. For example, in early-phase settings, directly test treatment efficacy after dose-finding (e.g. Wages & Tait, 2015); or in later phase trials directly continue to confirmatory stage (phase III) after learning stage (phase II) (e.g. Schmidli et al., 2006; Jiang & Yuan, 2023). In contrast, later-phase trials might incorporate adaptations based on more definitive efficacy or safety data.

1.2.2 Statistical Challenges and Literature Review of Two-Stage Seamless Adaptive Designs

Each category of the two-stage seamless adaptive design requires tailored statistical approaches to ensure the validity and integrity of trial results. The simplest category, the 0-D design, does not require major adjustments to standard statistical

methodology, as the objectives, endpoints, and patient populations remain consistent across both stages. In these cases, conventional inference procedures can be applied without introducing bias or compromising error control.

Historically, foundational work by Bauer and Köhne (1994) introduced a procedure for combining p-values into a single test statistic for adaptive interim analyses, laying the groundwork for maintaining type I error across seamless transitions. Posch et al. (2005) further developed flexible group sequential methods incorporating adaptive treatment selection while preserving the validity of both hypothesis testing and estimation, making their methodology especially relevant to seamless Phase II/III designs where treatment selection decisions are made at the interim.

However, as the number of differences increases, as seen in 1-D, 2-D, and 3-D designs, statistical complexity correspondingly grows. These designs introduce challenges in managing changes in trial objectives, study endpoints, and target populations between stages. Existing methods, particularly those originally developed for traditional group sequential or fixed-design trials, may not fully accommodate the adaptive features and dynamic changes present in seamless designs. When multiple elements such as the endpoint, objective, and population vary between stages, additional assumptions and methodological considerations are needed to derive valid statistical inferences. In such settings, approaches such as covariate adjustment, Bayesian frameworks, and advanced modelling strategies become necessary to maintain the rigor and reliability of results (see, e.g., Jiang & Yuan, 2023; Ma et al., 2022). In this section, we

review the primary statistical challenges and associated methods developed in the literature, organized by three major aspects of design divergence: (1) different objectives, (2) different study endpoints (including different endpoint data types), and (3) different patient populations.

1.2.2.1 Different Study Objectives

In scenarios where the study objectives differ between stages, such as transitioning from dose-finding in Phase II to confirmatory efficacy evaluation in Phase III, a dual-hypothesis framework is required., and statistical methods must ensure control of type I error rates across both stages. Early foundational work by Bauer and Köhne (1994) proposed combination tests that retain control of the familywise type I error rate while allowing mid-trial adaptation. Posch et al. (2005) extended this to allow for adaptive treatment selection in flexible group sequential settings. Their method is particularly relevant in seamless Phase II/III designs where early-stage treatments may be dropped or selected for continuation.

Jin and Zhang (2021) proposed a seamless 2-in-1 design that integrates Phase 2 and Phase 3 for treatment or dose selection, maintaining overall type I error control. This was later extended by Jin and Zhang (2023) to accommodate biomarker-based subpopulation selection, further increasing flexibility and clinical relevance. Stallard et al. (2014) and Bretz et al. (2006) proposed approaches to enable subgroup or hypothesis selection at interim, ensuring statistical rigor despite objective changes. Similarly, Schmidli et al. (2006) and Maca et al. (2006) discussed practical and regulatory aspects of confirmatory adaptive seamless designs that support changing objectives.

1.2.2.2 Different Study Endpoints

Endpoint divergence across stages, for example using a short-term surrogate marker in Stage 1 and a long-term clinical outcome in Stage 2, poses both conceptual and statistical challenges. In this type of scenario, it is usually assumed that the first-stage endpoint is associated with the second-stage endpoint, while predictiveness is asserted only when justified.

Chow, Lu, and Tse (2007) directly addressed this by proposing statistical methods to analyse two-stage seamless trials with different study endpoints, based on the assumption that the early endpoint is predictive of the final one. Takahashi et al. (2022) offered a bivariate binomial approach for combining short-term and long-term binary endpoints in a two-stage adaptive design. Their exact conditional test ensures robustness against dependence structure assumptions.

Chow and Lin (2015) discussed analysis methods for 2-D two-stage seamless adaptive designs in which the study objective and/or endpoint differ between stages. Their work focused on maintaining statistical validity when integrating interim and final analyses, particularly in settings involving transitions from surrogate marker to clinical outcomes.

Stallard (2010) introduced a framework for incorporating short-term endpoints into confirmatory decision-making, which was extended in Stallard et al. (2015) for flexible treatment selection. Kunz et al. (2014) compared previous methods for incorporating early outcomes, and proposed a new approach emphasizing power and probability of correct treatment selection. Friede et al. (2020) presented simulation-based

approaches for using early outcomes to guide treatment or subgroup selection and implemented these in R. The flexible simulation-based framework also supports mixed-type endpoints (e.g., binary, normal, time-to-events) by modelling early and final outcomes separately yet coherently in R.

In Bayesian contexts, Yang et al. (2024) introduced a Bayesian predictive power approach to jointly monitor multiple co-primary endpoints. Researchers have also developed specialized models for analysing time-to-event and count data under two-stage adaptive frameworks. Lu, Tse, and Chow (2010), Lu et al. (2012), and Lu, Chow, and Tse (2014) proposed inferential approaches for such data types under Weibull distribution assumptions, accounting for issues such as non-uniform entry, censoring, and differing study durations across stages.

Moreover, in therapeutic areas such as metabolic dysfunction-associated steatohepatitis (MASH), where disease progression is slow and clinical endpoints are difficult to capture early, adaptive trial designs have been proposed to facilitate development by enabling interim decisions based on surrogate measures (Filozof et al., 2017). This highlights the importance of statistical frameworks capable of integrating differing endpoints across trial stages, particularly in 2-D seamless designs.

1.2.2.3 Different Study Populations

In two-stage seamless adaptive designs, patient populations may shift between stages due to factors such as changes in eligibility criteria, biomarker-driven enrichment, emerging evidence on subgroup responsiveness, or progression in disease status. These

population shifts raise important concerns about internal validity, treatment effect generalizability, and the validity of pooled inference.

Rosenblum and Van der Laan (2011) proposed a design framework for optimizing the identification of beneficial subpopulations, defined by baseline covariates. Their method aims to maximize power to detect whether a treatment yields a positive effect in specific covariate-defined groups. While not framed explicitly for seamless adaptive trials, their use of weighted test statistics across subgroups offers conceptual tools for managing changing patient populations, especially in settings where subpopulation definitions evolve or interim analyses guide enrichment strategies.

To address the increasing interest in personalized medicine, Jin and Zhang (2023) developed an extension to their 2-in-1 adaptive design that allows for **biomarker subpopulation selection** at the interim. This makes it possible to refine the target population in Stage 2 while still preserving the inferential framework of a seamless design.

Jenkins et al. (2010) proposed an adaptive seamless Phase II/III design for oncology trials which allows evaluation of treatment effects in both the full population and a biomarker-defined subgroup as co-primary populations. Their method accounts for correlated survival endpoints and supports subpopulation selection at interim analysis, enabling inclusion or exclusion criteria to be updated based on early efficacy signals. This framework enables adaptive population refinement without compromising statistical validity.

Jiang and Yuan (2023) proposed four different seamless Phase II/III designs for dose optimization, comparing their efficiency and operating characteristics. While the designs aim to improve trial efficiency by selecting optimal doses early, the authors note that two of the four designs are particularly suitable when population changes across stages, offering greater robustness in the presence of population drift.

1.3 Research Gap and Problem Statement

Although two-stage seamless adaptive designs have gained increasing attention for their potential to improve trial efficiency and flexibility, important methodological challenges remain, particularly in complex applications such as biosimilar development. As clinical trials evolve toward more adaptive and data-driven frameworks, there is a pressing need for statistical methods that can support this flexibility without compromising scientific rigor. In particular, the integration of population shifts, changing endpoints, and adaptive decision-making criteria into a unified framework remains underdeveloped. This section outlines three specific areas where current methodologies fall short and where further research is needed to advance the effective use of two-stage seamless adaptive designs.

1.3.1 Shifts in Study Population and Change in Study Endpoints

An important but underexplored challenge in two-stage seamless adaptive designs is the scenario where both the study population and the primary endpoint change between stages. This situation is increasingly relevant in modern clinical trials, especially those targeting complex diseases or precision medicine settings, where disease may progress during the conduct of trial, and interim results often necessitate

change to the enrolled patient population and a transition from early surrogate markers to more clinically meaningful outcomes.

Despite the flexibility of adaptive designs, existing statistical methodologies typically accommodate only single-dimension adaptations, such as changes in the endpoint or the population, but not both. As reviewed in Section 1.2.2, most frameworks focus on controlling type I error and ensuring unbiased estimation under one-dimensional shifts. However, multi-dimensional adaptations, involving simultaneous changes to both the population and the endpoint, introduce analytical complexities that are not adequately addressed by current methods. These complexities may include correlated shifts in subgroup representation and response dynamics, or incompatibility of interim and final endpoints due to differing data distributions.

This gap presents a critical limitation for the design and analysis of adaptive seamless trials, as unadjusted shifts may lead to biased inference, reduced power, and compromised trial integrity. Addressing this methodological shortcoming requires the development of new statistical frameworks that jointly account for endpoint change and population shift, while preserving statistical validity of the trial analysis.

1.3.2 Challenges in Biosimilar Development

Biosimilars, the biologic products shown to be highly similar to approved reference biologics, pose distinct challenges in clinical development due to the complex nature of biological manufacturing and the high evidentiary standards required by regulatory agencies. The FDA's stepwise framework for biosimilar approval emphasizes a cumulative evidence approach, including analytical studies,

pharmacokinetic/pharmacodynamic (PK/PD) evaluations, and comparative clinical studies to confirm similarity in efficacy and safety. While this process ensures regulatory confidence, it is resource-intensive and often results in prolonged development timelines.

Given the high costs associated with biologics and the growing global demand for more affordable alternatives, there is increasing need to modernize biosimilar trial design. Conventional approaches, which treat each required study as independent, can lead to duplication of effort and delays in decision-making. This is particularly problematic in therapeutic areas such as cancer, autoimmune disorders, and metabolic diseases, where timely patient access to biosimilars could have a meaningful public health impact.

Despite growing interest in adaptive designs, their application in biosimilar development remains relatively limited. Few studies have systematically explored how flexible trial structures, such as two-stage seamless adaptive designs, can be effectively adapted to the unique regulatory and scientific requirements of biosimilar evaluation. Addressing this gap is essential for expanding the methodological toolkit available to support more efficient, rigorous, and evidence-based biosimilar approval processes.

1.3.3 Lack of Similarity Threshold Frameworks for Biosimilar Development

An essential element of adaptive clinical trial designs is the ability to make preplanned decisions at interim analyses, such as whether to continue, modify, or stop a trial based on accumulating evidence. These decisions are typically guided by statistical

decision boundaries, which serve as pre-specified thresholds to preserve type I error control and trial integrity.

While similarity threshold frameworks are well established for traditional group sequential and some adaptive designs, their extension to more complex settings, such as two-stage seamless trials for biosimilars, remains underdeveloped. Existing methods largely focus on superiority or efficacy trials, with limited applicability to biosimilarity objectives, which often involve equivalence margins and non-inferiority comparisons.

This gap is particularly critical in the context of biosimilar development when applying two-stage seamless adaptive designs, where interim decisions based on early pharmacokinetic (PK) data must be carefully evaluated and appropriately integrated with subsequent primary clinical endpoint analyses. In the absence of well-defined statistical frameworks to guide these decisions, the application of such designs remains largely ad hoc, increasing the risk of inconsistency with regulatory expectations and undermining the efficiency and credibility of the development process.

There is thus a critical need for research focused on developing similarity threshold methodologies specifically tailored for biosimilarity objectives within two-stage seamless designs. Advancing this area would support more consistent interim decision-making, improve trial efficiency, and enhance the acceptability of adaptive designs in regulatory review processes for biosimilars.

1.4 Research Objectives

The overarching goal of this thesis is to advance the statistical methodology for two-stage seamless adaptive designs, with a focus on improving their applicability in

complex and evolving clinical trial contexts. While these designs offer substantial operational and ethical advantages, their analytical foundations remain underdeveloped in scenarios involving changes to study populations, endpoints, or objectives across the stages, which are common in modern therapeutic areas such as oncology, liver disease, and biosimilar development. This research aims to address these methodological gaps through three interconnected objectives:

1.4.1 Development of a Statistical Method for Analysing Two-Stage Seamless Adaptive Designs with Population Shifts and Endpoint Changes

The first objective is to develop a robust statistical framework for analysing two-stage seamless adaptive trials in which both the patient population and the study endpoint change between stages. While most existing methods assume consistency in these design elements, real-world trials often face changes due to disease progression, evolving eligibility criteria, or shifts from surrogate marker to clinical outcomes.

This research proposes a method capable of integrating data across such shifts while maintaining statistical validity, particularly focusing on preserving type I error and ensuring adequate power. The framework will be applicable to areas such as liver disease, where dynamic patient characteristics and evolving endpoints are common. Its performance is evaluated through simulation studies designed to assess power and robustness under various scenarios.

1.4.2 Proposal for the Use of a Two-Stage Seamless Adaptive Design in Biosimilar Product Development

The second objective is to design a novel two-stage seamless adaptive trial structure tailored to the unique regulatory and scientific requirements of biosimilar drug

development. The current stepwise approach, which typically separating pharmacokinetic (PK) and clinical evaluations, is often inefficient and costly. This research proposes integrating both phases within a unified design that allows for early decisions based on PK findings at the end of stage 1.

A key innovation is the incorporation of a biosimilarity index derived from reproducibility probability, which supports a continuous assessment of similarity while aligning with regulatory expectations for biosimilarity. The design aims to reduce redundancy, enable early conclusions when justified, and provide a more efficient alternative to traditional biosimilar pathways.

1.4.3 Extension: Development of a Similarity Threshold Framework Using the Relative Biosimilarity Index in Biosimilar Product Development

Building on the two-stage adaptive design framework proposed in Section 1.4.2, the third objective of this research is to extend the methodology by introducing a new statistical index—the Relative Biosimilarity Index (RBI)—and developing a corresponding framework for similarity threshold specification. While the earlier design incorporates a biosimilarity index to facilitate early decision-making, this extension focuses on using RBI to formally test biosimilarity and guide adaptive decisions at interim analyses.

The RBI is constructed as a ratio between the observed biosimilarity index and the reference-based reproducibility probability, offering a standardized and interpretable metric for assessing biosimilarity under uncertainty. This objective aims to provide guidance on appropriate thresholds selection for RBI to determine whether a

biosimilar candidate can be considered sufficiently similar at the interim stage or whether further confirmatory data are required.

The proposed framework will include methods for calibrating the similarity threshold to control type I error while achieving desirable power. Simulation studies are conducted to evaluate the statistical performance of the RBI-based boundary under various trial conditions. This work directly supports the broader goal of improving decision-making in seamless adaptive biosimilar trials by providing a scientifically grounded and regulatorily aligned approach to interim evaluation.

1.5 Significance of the Research

This research contributes to the advancement of two-stage seamless adaptive trial methodology by addressing key analytical limitations that currently hinder its broader application in complex and evolving clinical settings. While the advantages of these designs, such as improved efficiency, ethical responsiveness, and accelerated decision-making, are increasingly recognized, their practical implementation remains constrained in scenarios involving changes to patient populations and study endpoints across the stages. These challenges are especially relevant in therapeutic areas marked by disease progression, as well as in biosimilar development, where trial conduct is closely governed shaped by regulatory expectations.

A central contribution of this work is the development of an analytical framework that enables valid statistical inference when both the study population and endpoint are different between stages. This reflects real-world complexity, particularly in areas like liver disease, where early surrogate markers are often used and patient

characteristics may shift over time. The proposed method aims to improve the adaptability and robustness of statistical analysis under these conditions, supporting reliable conclusions without compromising statistical integrity.

In the context of biosimilar studies, this research introduces a two-stage design that integrates pharmacokinetic and comparative clinical evaluation phases, offering a streamlined alternative to the traditional stepwise approach. By incorporating a reproducibility-based biosimilarity index and extending it to a testable Relative Biosimilarity Index (RBI), this work provides a statistically valid and interpretable framework for assessing biosimilarity in a manner consistent with analytical-similarity principles and regulatory standards (Chow, 2018) and conducive to more timely decision-making. Finally, this work introduces the Relative Biosimilarity Index (RBI), which extends BI by explicitly incorporating inherent variability and providing a method to define a similarity threshold that is interpretable across products, thereby improving transparency and helping to standardise review and approval processes; this aligns with FDA's BsUFA III regulatory science focus on streamlining evidentiary packages for biosimilarity (FDA, 2024; FDA, 2025).

In summary, this research bridges the methodological gaps in two-stage seamless adaptive trial design and introduces practical innovations for managing population and endpoint heterogeneity, supporting biosimilar development, and guiding interim decision processes. These contributions are expected to advance the field of adaptive design and promote the timely development of accessible and cost-effective therapies.

1.6 Thesis Organization

This thesis is organized into five chapters, each contributing to the overarching goal of advancing statistical methodologies for two-stage seamless adaptive clinical trial designs.

Chapter 1 introduces the background and motivation for the study, with a focus on the literature review and the limitations of current adaptive trial methodologies when applied to complex scenarios involving changes in study populations and endpoints. It also outlines specific challenges in biosimilar development and the lack of rigorous similarity threshold frameworks. The chapter concludes with clearly defined research objectives and a discussion of the significance of the proposed work.

Chapter 2 presents statistical methods for analysing two-stage seamless adaptive designs when both the study population and endpoint differ between stages. It addresses the analytical complexity that arises in such settings and proposes a framework that ensures valid inference. The methods are motivated by clinical contexts where endpoint switching and population heterogeneity are common, such as liver disease.

Chapter 3 proposes a novel two-stage adaptive trial design specifically tailored to biosimilar development. It integrates pharmacokinetic and comparative clinical studies into a single cohesive framework, aiming to improve efficiency while meeting regulatory standards. A key feature of this design is the application and evaluation on a biosimilarity index, which supports earlier and more informed decision-making.

Chapter 4 extends this work by developing a statistical framework for determining biosimilarity using the Relative Biosimilarity Index (RBI). This chapter focuses on constructing principled testing procedures and thresholds for decisions in biosimilar-focused adaptive trials, ensuring both operational efficiency and statistical rigor.

Chapter 5 concludes the thesis by summarizing the key contributions and discussing their implications for adaptive trial design and regulatory science. It also outlines directions for future research, including potential extensions to other data types and therapeutic areas.

Together, these chapters form a cohesive body of work that addresses critical methodological gaps in two-stage seamless adaptive designs and contributes practical innovations applicable to modern clinical research, particularly in biosimilar development.

2. Analysis of Innovative Two-stage Seamless Adaptive Design with Different Endpoints and Population Shift

In this chapter, building on the background and research objectives outlined in Chapter 1, this chapter corresponds to the topic regarding the analysis of two-stage seamless adaptive designs with different endpoints when there is a population shift (Mai & Chow, 2024). Mai and Chow (2024) addressed one of the key methodological gaps identified earlier: the lack of valid statistical methods when both study endpoints and patient populations change across stages in a seamless adaptive design. By developing an approach that predicts primary endpoints for interim analyses and combines information before and after a population shift, this work provides a rigorous framework for valid inference, type I error control, and robust power in such complex adaptive trial setting.

2.1 Introduction

In recent years, the use of seamless adaptive design in clinical trials for evaluation of safety and efficacy of a test treatment under investigation has become very popular in clinical research and development (FDA, 2019). Adaptive trial design provides the flexibility, efficiency, and opportunity to make changes during the conduct of the clinical trials based on accumulated data observed at the interim (FDA, 2019). A

seamless design is a study design that combines several separate (independent) trials into a single study (Chow & Chang, 2011). Such a seamless design is able to address objectives of individual studies with similar or different endpoints. A seamless adaptive trial design allows adaptations (modifications) to the study protocol during the conduct of trial after the review of the accumulated data observed at interim. All data collected from enrolled patients before and after the adaptation would be included in the final analysis. Such a design cannot only reduce the lead time between individual studies but also expedite the development process. Most importantly, it is cost-effective and increase the probability of success of drug development. For example, in a typical two-stage seamless adaptive trial design, adaptations as recommended by an independent data safety monitoring committee (IDMC) based on the review of accrued data at interim or at the end of exploratory stage are allowed (Chow & Lin, 2015), which provides the opportunity for stopping trial early due to safety and/or futility/efficacy.

With the increasing applications of two-stage design, two-stage adaptive design and two-stage seamless adaptive design, many studies have been conducted to provide reliable and efficient statistical methods for the analysis. One of the most commonly used method was Simon's two-stage design, which allows the early stopping due to only futility, which is relatively conservative. On the basis of this, some researchers such as Thompson and Mander (2010) have extended the design to allow early stopping for futility. However, these designs did not pay much attention to the case when the study endpoints and/or study population are different. For the change in study population, a

study focusing on the population shift due to protocol amendments was conducted by Chow and Shao (2005) and a relevant method of analysis was proposed.

Nonetheless, when considering the two-stage seamless adaptive design, these methods may not be directly applicable. Chow, Lu and Tse (2007) has discussed the method for analysing seamless trial when the study endpoints of two stages are different. Chow and Lin (2015) have then derived and discussed statistical methods for analysis of the four different types of two-stage seamless adaptive designs they classified (study population are assumed to be unchanged). Fan, Zhao and Li (2020) have also proposed a 2-in-1 adaptive phase 2/3 designs which allows early decision based on intermediate endpoint. While these methods allow the change in study endpoints, the analysis method when there is a shift in patient population, which may be due to the relax of eligibility criteria and/or disease progression, was rarely mentioned.

During many two-stage seamless adaptive clinical trials, there is usually a need to adjust the study population through protocol amendments. Moreover, with the increasing prevalence of liver-related morbidity and mortality caused by certain liver diseases the development of treatment for liver diseases such as hepatitis B virus (HBV), hepatitis C virus (HCV), Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) and Metabolic Dysfunction-Associated Steatohepatitis (MASH) have thereby received much attention. Relevant therapeutic interventions are therefore in much need, while during the clinical trials for these medical treatments, disease progression is often

observed. Therefore, there is an urgent need to have a valid statistical method for making reliable analyses of these trials.

The objective of this study is to develop a statistical approach for analysing two-stage seamless adaptive designs in which both the study population and the study endpoints change across stages. To this end, Section 2.2 provides relevant background, and Section 2.3 introduces the proposed analysis method that jointly accounts for endpoint differences and population shifts. Section 2.4 discusses type I error control, while Section 2.5 summarizes sample size considerations based on stage 2. The performance and robustness of the proposed method are evaluated through simulation studies in Section 2.6. Finally, concluding remarks are presented in Section 2.7.

2.2 Background and Motivation

As summarized in Chapter 1, two-stage seamless adaptive designs integrate exploratory and confirmatory stages within a single protocol and can be classified into categories depending on whether study objectives, endpoints, or patient populations differ between stages (Chow & Lin, 2015; Chow, 2020). While existing methodologies address certain settings such as designs with consistent populations or with endpoint changes alone, the case where both the study population and the study endpoints differ across stages has received limited attention. This chapter is motivated by that gap and develops statistical methods for valid inference under these more complex scenarios, with a particular focus on ensuring type I error control and maintaining power in the presence of endpoint switching and population shifts.

2.3 Method of Data Combination when both Shift in Population and Endpoint Difference Occur

Without loss of generality and for illustration purpose, in this study we assume that the shift in population is due to the protocol amendment on the inclusion criteria. The idea can be similarly applied to the situation where the population shift is due to disease progression during the conduct of the clinical trial. The mapping from pre-shift to post-shift means is fitted using both pre- and post-shift outcomes (shared estimation), which makes the two Stage-2 estimators correlated within each arm.

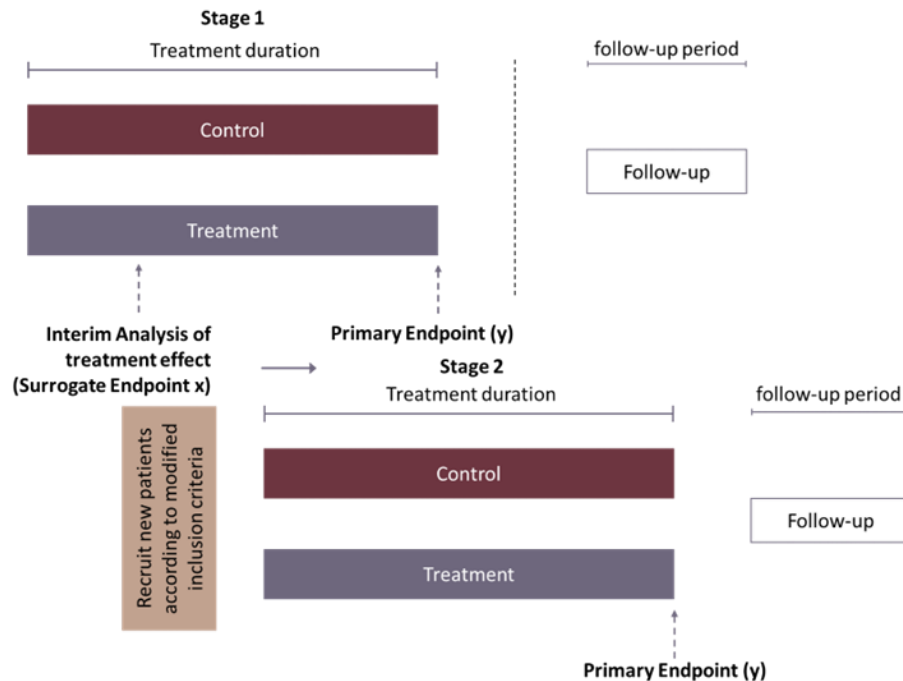


Figure 1: Illustration of two-stage seamless design. Adapted from Mai, W., & Chow, S.-C. (2024). Journal of Biopharmaceutical Statistics.

2.3.1 Step 1: Interim Analysis at the End of Stage 1 (before protocol amendment)

First consider the group of patients who were recruited at the beginning of stage

1. Denote the endpoint data in stage 1 as $x_i, i = 1, \dots, 2n_1$, which is normally distributed

with mean μ_x and variance σ_x^2 . Following the idea by Chow, Lu and Tse (2007), suppose the primary endpoint data collected in stage 2 can be denoted as y_i , and there exists a known relationship between x and y such that

$$y_i = a + bx_i + \epsilon, \quad \mathbb{E}(\epsilon_i|x_i) = 0, \quad \text{Var}(\epsilon_i) = \sigma_\epsilon^2.$$

Since this is a two-stage trial, the overall objective and hypothesis can be expressed as the intersection of the two individual hypothesis tests at the interim analysis and the final analysis, i.e., $H_0: H_{01} \cap H_{02}$, where H_{01} and H_{02} are the null hypotheses at the two analyses respectively. In order to make an overall interpretation, the endpoints tested at each analysis should be the same, therefore, we can use the surrogate marker data collected in stage 1 to make predictions about the primary endpoint and denote these predictions as \hat{y}_{1i} . Then we can use this as the endpoint for interim analysis at the end of stage 1.

In order to make interim analysis at stage 1 using the primary clinical endpoint, we estimate the primary endpoint data with the above relationship and the collected data for surrogate marker.

$$\hat{y}_{1i} = a + bx_i$$

It is known that \bar{x}_j is the unbiased estimate of $\mu_{x,j}$, and $\bar{\hat{y}}_{1,j}$ is therefore the unbiased estimates of $a + b\mu_{x,j}$ for arm $j \in \{T, C\}$. We can then use the data to make unbiased estimates of the population variances of each treatment group:

$$S_{x,j}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x}_j)^2, \quad S_{\hat{y}_{1,j}}^2 = b^2 S_{x,j}^2,$$

$$\widehat{\text{Var}}(\bar{\hat{y}}_{1,j}) = \frac{b^2 S_{x,j}^2}{n_1} = \frac{b^2 S_{\hat{y}_{1,j}}^2}{n_1}.$$

We will then conduct the interim analysis using the predicted primary endpoint data, and the test statistics with respect to each type of hypothesis are illustrated in the following table:

Table 3: Statistical tests for various hypotheses testing at stage 1.

	Hypotheses	Test Statistic
Equality	$H_0: \mu_C = \mu_T$ $H_1: \mu_C \neq \mu_T$	$T_1 = \frac{\bar{y}_{1,T} - \bar{y}_{1,C}}{\sqrt{\widehat{Var}(\bar{y}_{1,T}) + \widehat{Var}(\bar{y}_{1,C})}}$
Superiority	$H_0: \mu_T - \mu_C \leq \delta$ $H_1: \mu_T - \mu_C > \delta$	$T_1 = \frac{\bar{y}_{1,T} - \bar{y}_{1,C} - \delta}{\sqrt{\widehat{Var}(\bar{y}_{1,T}) + \widehat{Var}(\bar{y}_{1,C})}}$
Non-inferiority	$H_0: \mu_T - \mu_C \leq -\delta$ $H_1: \mu_T - \mu_C > -\delta$	$T_1 = \frac{\bar{y}_{1,T} - \bar{y}_{1,C} + \delta}{\sqrt{\widehat{Var}(\bar{y}_{1,T}) + \widehat{Var}(\bar{y}_{1,C})}}$
Equivalence	$H_0: \mu_T - \mu_C \geq \delta$ $H_1: \mu_T - \mu_C < \delta$	Simply construct the confidence interval for $\mu_T - \mu_C$ using $\bar{y}_{1,T} - \bar{y}_{1,C}$.

The value p_1 is the p -value calculated from selected test above. In stage 1, since we allow the early stopping for efficacy or futility, we can give a range to p_1 for continuing the trial into the second stage:

Table 4: Decision boundaries at the end of stage 1.

Continue	$\alpha_1 < p_1 < \beta_1$
Stop for futility	$p_1 \geq \beta_1$
Stop for efficacy	$p_1 \leq \alpha_1$

2.3.2 Step 2: Collect Information at the End of Stage 2

For arm j , primary endpoint means before and after the population shift:

$$\{y_{1i}\}_{i=1}^{n_1} \sim (\mu_{y_{1,j}}, \sigma_{y_{1,j}}^2), \quad \{y_{2i}\}_{i=1}^{n_2} \sim (\mu_{y_{2,j}}, \sigma_{y_{2,j}}^2),$$

with sample means \bar{y}_{1j} (pre-shift) and \bar{y}_{2j} (post-shift), variances $Var(\bar{y}_{1j}) = S_{1j}^2/n_1$, $Var(\bar{y}_{2j}) = S_{2j}^2/n_2$, and $Cov(\bar{y}_{1j}, \bar{y}_{2j}) = 0$ (disjoint cohorts). Then, we can use the table below to summarise the information we have at the end of stage 2 for the final analysis on treatment effect.

Table 5: Information collected at each stage.

	Stage 1	Stage 2
Objective	Efficacy	Efficacy
Endpoint	Surrogate Marker	Primary Clinical Endpoint
Sample size	$2n_1$	$2n_1 + 2n_2$
Population	Only data from population before shift	Data from both population before and after shift

2.3.3 Step 3: Consider Population Shift (mean-level model, per arm)

Mean-level model (design-level covariates $R_k \in \mathbb{R}^p$):

$$\mu_{k,j} = \beta_{0j} + \beta_j^\top R_k, \quad k \in \{a, b\}, \quad \Delta R := R_a - R_b, \quad d_j := \mu_{a,j} - \mu_{b,j} = \beta_j^\top \Delta R.$$

Arm-specific WLS using pre & post means. Define

$$X_j = \begin{bmatrix} 1 & R_b^\top \\ 1 & R_a^\top \end{bmatrix}, \quad \bar{y}_j = \begin{bmatrix} \bar{y}_{1j} \\ \bar{y}_{2j} \end{bmatrix}, \quad W_j = \text{diag}(w_{1j}, w_{2j}),$$

with $w_{1j} = \widehat{Var}(\bar{y}_{1j})^{-1}$, $w_{2j} = \widehat{Var}(\bar{y}_{2j})^{-1}$. Then

$$\hat{\beta}_j = (X_j^\top W_j X_j)^{-1} X_j^\top W_j \bar{y}_j, \quad \hat{d}_j = \Delta R^\top \hat{\beta}_j.$$

Let

$$A_j := \Delta R^\top (X_j^\top W_j X_j)^{-1} X_j^\top W_j = [c_{j1} \quad c_{j2}] \Rightarrow \hat{d}_j = c_{j1} \bar{y}_{1j} + c_{j2} \bar{y}_{2j}.$$

Linear representation of the post-shift mean predicted from pre-shift:

$$\theta_{2j} = \bar{y}_{1j} + \hat{d}_j = (1 + c_{j1})\bar{y}_{1j} + c_{j2}\bar{y}_{2j} =: a_{1j}\bar{y}_{1j} + a_{2j}\bar{y}_{2j},$$

so

$$Var(\theta_{2j}) = a_{1j}^2 Var(\bar{y}_{1j}) + a_{2j}^2 Var(\bar{y}_{2j}), \quad Cov(\theta_{2j}, \bar{y}_{2j}) = a_{2j} Var(\bar{y}_{2j}).$$

Correlation-aware combination:

$$\hat{\mu}_{a,j} = \omega_j^* \bar{y}_{2j} + (1 - \omega_j^*) \theta_{2j},$$

$$\omega_j^* = \frac{Var(\theta_{2j}) - Cov(\theta_{2j}, \bar{y}_{2j})}{Var(\theta_{2j}) + Var(\bar{y}_{2j}) - 2Cov(\theta_{2j}, \bar{y}_{2j})},$$

$$Var(\hat{\mu}_{a,j}) = (\omega_j^*)^2 Var(\bar{y}_{2j}) + (1 - \omega_j^*)^2 Var(\theta_{2j}) + 2\omega_j^*(1 - \omega_j^*)Cov(\theta_{2j}, \bar{y}_{2j}).$$

2.3.4 Step 4: Hypothesis Testing at Stage 2 (post-shift treatment effect)

As the main interest is the treatment effect after population shift at the end of stage 2 of the clinical trial, hypothesis testing is conducted only on μ_a . Use the arm-wise $\hat{\mu}_{a,j}$ above and test the contrast

$$\hat{\Delta}_2 = \hat{\mu}_{a,T} - \hat{\mu}_{a,C}, \quad \widehat{Var}(\hat{\Delta}_2) = Var(\hat{\mu}_{a,T}) + Var(\hat{\mu}_{a,C}) \text{ (arms independent).}$$

Then we can have the relevant test statistics for each type of hypothesis:

Table 6: Statistical tests for various hypotheses testing at stage 2.

	Hypotheses	Test Statistic
Equality	$H_0: \mu_{a,C} = \mu_{a,T}$ $H_1: \mu_{a,C} \neq \mu_{a,T}$	$T_2 = \frac{\hat{\mu}_{a,T} - \hat{\mu}_{a,C}}{\sqrt{\widehat{Var}(\hat{\mu}_{a,T}) + \widehat{Var}(\hat{\mu}_{a,C})}}$
Superiority	$H_0: \mu_{a,T} - \mu_{a,C} \leq \delta$ $H_1: \mu_{a,T} - \mu_{a,C} > \delta$	$T_2 = \frac{(\hat{\mu}_{a,T} - \hat{\mu}_{a,C}) - \delta}{\sqrt{\widehat{Var}(\hat{\mu}_{a,T}) + \widehat{Var}(\hat{\mu}_{a,C})}}$

Non-inferiority	$H_0: \mu_{a,T} - \mu_{a,C} \leq -\delta$ $H_1: \mu_{a,T} - \mu_{a,C} > -\delta$	$T_2 = \frac{(\hat{\mu}_{a,T} - \hat{\mu}_{a,C}) + \delta}{\sqrt{\widehat{Var}(\hat{\mu}_{a,T}) + \widehat{Var}(\hat{\mu}_{a,C})}}$
Equivalence	$H_0: \mu_{a,T} - \mu_{a,C} \geq \delta$ $H_1: \mu_{a,T} - \mu_{a,C} < \delta$	Simply construct the confidence interval of $\mu_{a,T} - \mu_{a,C}$

Without loss of generality, the test for equality is discussed here. Since all the data points are assumed to be independently, identically normally distributed, both

$$\frac{\hat{\mu}_{a,T} - \mu_{a,T}}{\sqrt{\widehat{Var}(\hat{\mu}_{a,T})}} \quad \text{and} \quad \frac{\hat{\mu}_{a,C} - \mu_{a,C}}{\sqrt{\widehat{Var}(\hat{\mu}_{a,C})}}$$

converge in distribution to a standard normal distribution. Therefore, it is reasonable to claim that under the null hypothesis, the test statistic

$$T = \frac{\hat{\mu}_{a,T} - \hat{\mu}_{a,C}}{\sqrt{\widehat{Var}(\hat{\mu}_{a,T}) + \widehat{Var}(\hat{\mu}_{a,C})}}$$

has a limiting standard normal distribution according to the Slutsky's Theorem.

The corresponding p-value (p_2) can then be calculated, and to maintain overall type I error control, it is compared against the critical value α_2 . A detailed discussion of the derivation of α_2 is provided in Section 2.4.

Table 7: Decision boundaries at the end of stage 2.

Stop for biosimilar	$p_2 \leq \alpha_2$
Stop for not biosimilar	$p_2 > \alpha_2$

2.4 Type I Error Control

2.4.1 Setup and Notation

Let p_1 and p_2 be the p-values from the tests at Stage 1 and Stage 2. Define the corresponding Z-scores

$$T_k = \Phi^{-1}(1 - p_k), \quad k \in \{1, 2\},$$

so that larger T_k favours biosimilarity. Stage 1 proceeds to Stage 2 only if

$$\mathcal{P} = \{z_{F_1} < T_1 < z_{E_1}\}, \quad z_{E_1} = \Phi^{-1}(1 - \alpha_1), \quad z_{F_1} = \Phi^{-1}(1 - \beta_1).$$

Let the familywise error target be α and define the remaining after Stage 1 as

$$\alpha_{rem} = \alpha - \Pr_{H_0}(p_1 \leq \alpha_1).$$

From exposition, when p_1 is (approximately) uniform under H_0 , $\Pr_{H_0}(p_1 \leq \alpha_1) \approx \alpha_1$ and

$$\Pr_{H_0}(\mathcal{P}) \approx \beta_1 - \alpha_1.$$

Conditional-Error Rule

Under H_0 , the joint behaviour of (T_1, T_2) is well approximated by a bivariate normal model

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad -1 \leq \rho \leq 1,$$

where ρ quantifies the stage-to-stage correlation (e.g., due to data reuse or shared nuisance estimation). Conditioning on the observed interim value $T_1 = t$ and on proceeding, the Stage-2 distribution is

$$T_2 | T_1 = t \sim \mathcal{N}(\rho t, 1 - \rho^2).$$

Spending the pre-proceed budget

$$q = \frac{\alpha_{rem}}{\Pr_{H_0}(\mathcal{P})} \approx \frac{\alpha - \alpha_1}{\beta_1 - \alpha_1}.$$

via the moving boundary

$$c_2(T_1) = \rho T_1 + \sqrt{1 - \rho^2} z_{1-q} \Leftrightarrow p_2 \leq 1 - \Phi\left(\rho T_1 + \sqrt{1 - \rho^2} z_{1-q}\right)$$

(reject at Stage 2 if $T_2 \geq c_2(T_1)$) ensures

$$\Pr_{H_0}(\text{Stage 2 reject}, \mathcal{P}) = q \Pr_{H_0}(\mathcal{P}) = \alpha_{rem}$$

and hence overall type I error α . When $\rho = 0$, $c_2(T_1) = z_{1-q}$ (constant threshold).

2.4.2 Estimating ρ in Practice

2.4.2.1 Design-Based plug-in (when Stage 2 reuses Stage 1)

Use the predicted-primary interim contrast

$$\Delta_1^{(pred)} = \bar{y}_{1,T} - \bar{y}_{1,C} = b(\bar{x}_T - \bar{x}_C), \quad Var(\Delta_1^{(pred)}) = b^2 \left(\frac{S_{x,T}^2}{n_1} + \frac{S_{x,C}^2}{n_1} \right).$$

With the stage 2 final contrast

$$\Delta_2^{(comb)} = \hat{\mu}_{a,T} - \hat{\mu}_{a,C}, \quad Var(\Delta_2^{(comb)}) = Var(\hat{\mu}_{a,T}) + Var(\hat{\mu}_{a,C}).$$

Under the relationship $y = a + bx + \epsilon$ and independence of ϵ and x ,

$$Cov(\bar{x}_j, \bar{y}_{1j}) = b Var(\bar{x}_j) = b \frac{S_{x,j}^2}{n_1},$$

and $Cov(\bar{x}_j, \bar{y}_{2j}) = 0$ (different cohort).

Therefore,

$$Cov(\Delta_1^{(pred)}, \Delta_2^{(comb)}) = b^2 \left[(1 - \omega_T^*) a_{1T} \frac{S_{x,T}^2}{n_1} + (1 - \omega_C^*) a_{1C} \frac{S_{x,C}^2}{n_1} \right],$$

and finally,

$$\hat{\rho} \approx \frac{Cov(\Delta_1^{(pred)}, \Delta_2^{(comb)})}{\sqrt{Var(\Delta_1^{(pred)}) Var(\Delta_2^{(comb)})}}$$

2.4.2.2 Pilot-based Monte Carlo

Simulate the planned design under H_0 with planned (n_1, n_2) , relationship coefficients (a, b) , WLS step (thus $a_{1j}, a_{2j}, \omega_j^*$), and decision rules. For each replicate compute

$$T_1 = \Phi^{-1}(1 - p_1), \quad T_2 = \Phi^{-1}(1 - p_2)$$

then set $\hat{\rho} = Cor(T_1, T_2)$ and use this in $c_2(T_1)$.

2.5 Sample Size Calculation

Under the alternative hypothesis, compute the required sample size. Assuming simple randomisation is used, and randomisation ratio is 1:1, then we can summarise the sample size calculation in the table below.

Table 8: Sample size requirement for various hypotheses testing.

	Hypotheses	Sample Sizes
Equality	$H_0: \mu_{2,C} = \mu_{2,T}$ $H_1: \mu_{2,C} \neq \mu_{2,T}$	$n_{2,total} = \frac{4 \left(\frac{z_{\alpha_2}}{2} + z_{\beta} \right)^2 \sigma_2^2}{(\mu_{2,T} - \mu_{2,C})^2}$
Superiority	$H_0: \mu_{2,T} - \mu_{2,C} \leq \delta$ $H_1: \mu_{2,T} - \mu_{2,C} > \delta$	$n_{2,total} = \frac{4(z_{\alpha_2} + z_{\beta})^2 \sigma_2^2}{(\mu_{2,T} - \mu_{2,C} - \delta)^2}$
Non-inferiority	$H_0: \mu_{2,T} - \mu_{2,C} \leq -\delta$ $H_1: \mu_{2,T} - \mu_{2,C} > -\delta$	$n_{2,total} = \frac{4(z_{\alpha_2} + z_{\beta})^2 \sigma_2^2}{(\mu_{2,T} - \mu_{2,C} + \delta)^2}$
Equivalence	$H_0: \mu_{2,T} - \mu_{2,C} \geq \delta$ $H_1: \mu_{2,T} - \mu_{2,C} < \delta$	$n_{2,total} = \frac{4(z_{\alpha_2} + z_{\beta})^2 \sigma_2^2}{(\delta - \mu_{2,T} - \mu_{2,C})^2}$

Here $\mu_{2,T} - \mu_{2,C}$ is the difference that the trial wants to detect at α_2 significance level and with power $1 - \beta$.

In the next section, the relationship between the sample size allocation ratio between the two stages and the overall power of the trial is studied via simulation (with type I error controlled).

2.6 Simulation Study and Sensitivity Analysis

In this section, simulation study on the performance of the proposed method with different allocation ratio is discussed. In order to test the robustness of the performance, the parameters such as sample size, β_1 and effect size are modified. The relationship between surrogate marker and primary endpoint is set to be $y = 4 + 6x$.

2.6.1 $\beta_1 < \alpha_2$

Suppose the effect size is 2, the means of surrogate marker used at the interim analysis for the treatment and control group are 0.4 and 0.6 respectively. Then the standard deviations are thereby assumed to be both 0.1. The means of primary endpoint for the treatment and control group are 6.4 and 7.6 for the population before shift, and 7.4 and 8.6 for the population after shift. The standard deviations of both treatments are 0.6 and 0.6 respectively for the populations before and after shift. Without loss of generality, in the first set of simulation following assumptions were also made: $\alpha_1 = 0.01$, $\beta_1 = 0.35$, $\alpha = 0.025$, the corresponding α_2 for stage 2 can thereby be calculated.

Figures 2-4 are generated from 5000 simulations with sample size per arm 100, 50 and 30 respectively using R. From the plots below, we can easily identify the appropriate allocation ratios between two stages ($r = n_1/(n_1 + n_2)$) with a desired statistical power such as 80% can be achieved. Moreover, we can conclude that the power of the whole analysis increases with the increase in allocation ratio, i.e., with the same total number of patients, the more patients recruited in stage 1, the higher the power would be.

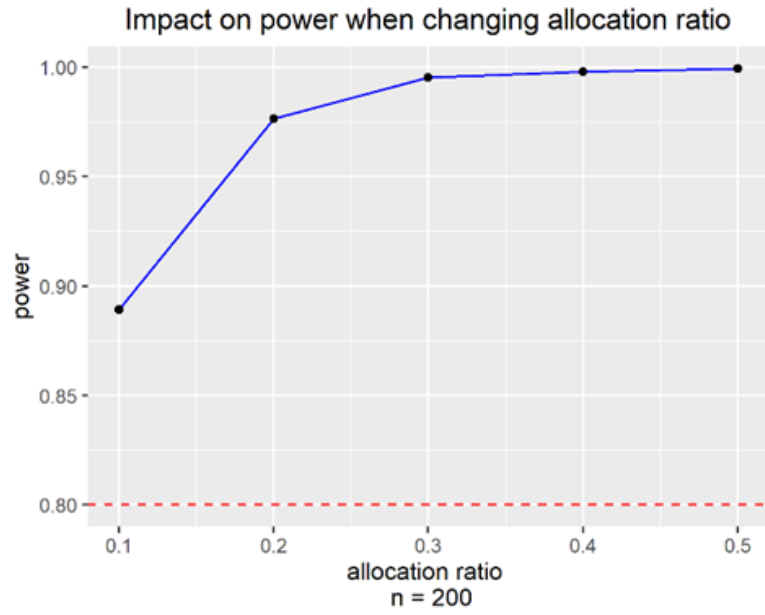


Figure 2: Simulation results with sample size per arm 100.

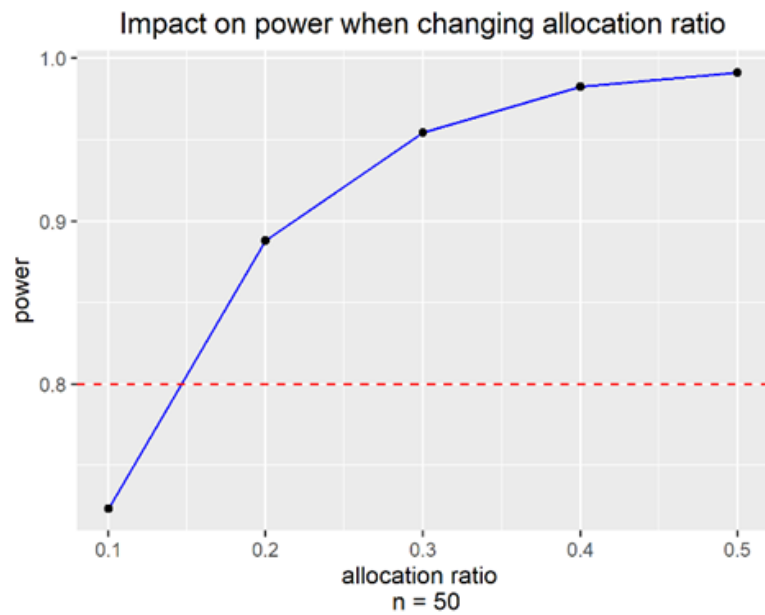


Figure 3: Simulation results with sample size per arm 50.

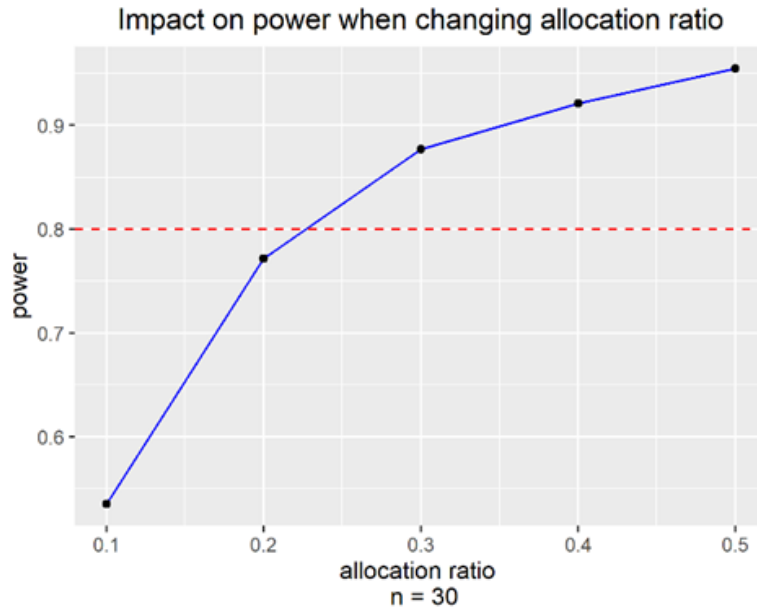


Figure 4: Simulation results with sample size per arm 30.

Table 9: Power with different allocation ratio when sample size per arm is 100.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.889	0.977	0.995	0.998	0.999

Table 10: Power with different allocation ratio when sample size per arm is 50.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.723	0.888	0.954	0.982	0.991

Table 11: Power with different allocation ratio when sample size per arm is 30.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.535	0.772	0.877	0.921	0.955

From the trends shown in the plots and tables, larger sample sizes can lead to higher statistical power and therefore requires smaller allocation ratio to reach the desired power for final analysis.

2.6.2 $\beta_1 \geq \alpha_2$

When all other settings remain unchanged, by assuming that $\beta_1 < \alpha_2$, the validity of the proposed method can be tested under different assumption of β_1 . Without loss of generality, let $\beta_1 = 0.15$ and the corresponding α_2 for stage 2 is computed.

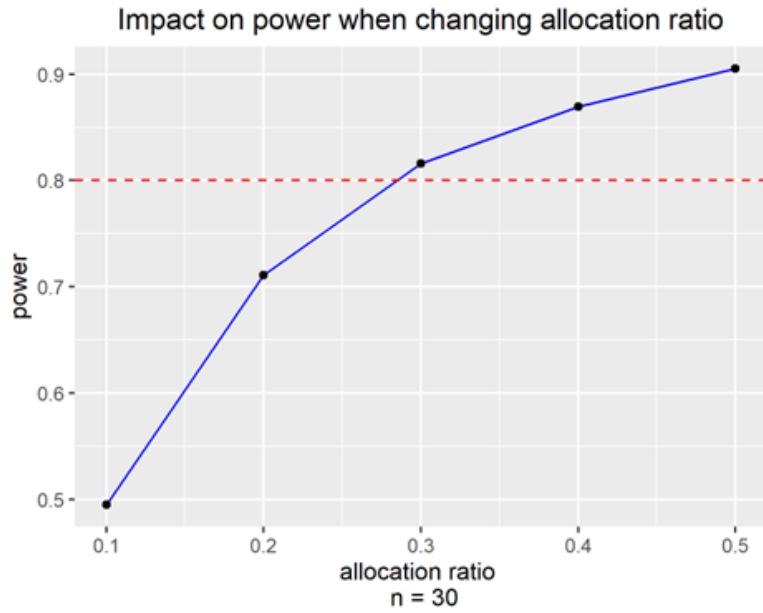


Figure 5: Simulation results with sample size per arm 30.

Table 12: Power with different allocation ratio when sample size per arm is 30.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.495	0.711	0.816	0.869	0.905

From this simulation, it can be identified that when $\beta_1 < \alpha_2$, the power of the test is slightly different from that when $\beta_1 \geq \alpha_2$, but by choosing the appropriate allocation ratio, the desired power can still be achieved.

2.6.3 Different Effect Size

With the same assumptions of settings as in Section 6a, now let the effect size be 1.5 and sample size per arm be 100. The means of surrogate marker and primary endpoint remain unchanged, while the standard deviations change accordingly.

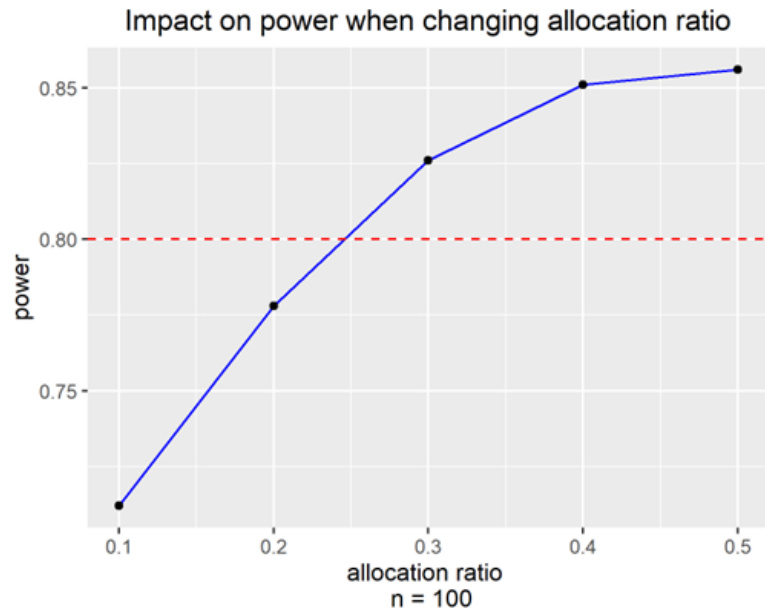


Figure 6: Power change when sample size per arm 100, effect size 1.5.

Table 13: Power with different allocation ratio when sample size per arm is 100, effect size 1.5.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.712	0.778	0.826	0.851	0.856

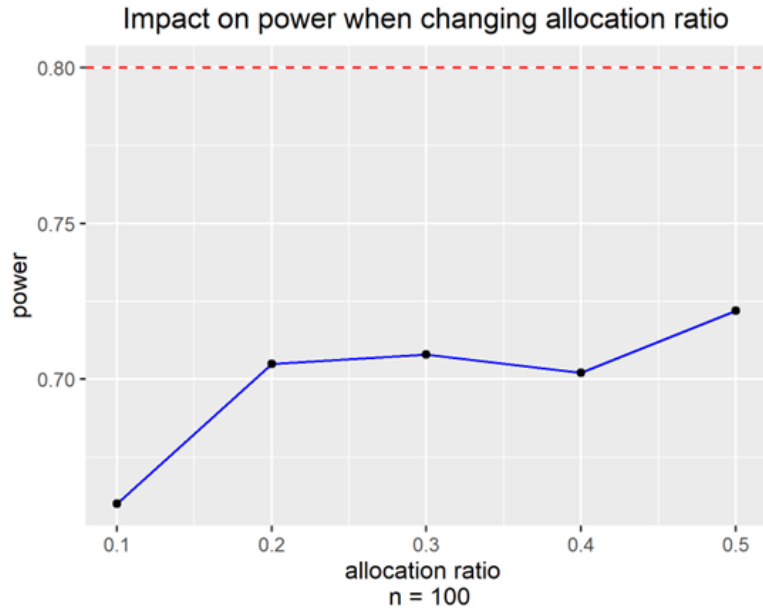


Figure 7: Power change when sample size per arm 100, effect size 1.4.

Table 14: Power with different allocation ratio when sample size per arm is 100, effect size 1.4.

Ratio	0.1	0.2	0.3	0.4	0.5
Power	0.660	0.705	0.708	0.702	0.722

By changing the effect size and comparing the result, it is clear that the smaller the effect size, the smaller the power, which coincide with common anticipation, since the treatment effect is harder to be identified. It should be noted that when the effect size is equal to 1.4, the power will be much lower than the normally desired value, i.e. 80%. Therefore, the proposed method should have better performance when the effect size is larger than 1.4.

From the plots and tables above, it is reasonable to say that the proposed method has satisfiable performance in terms of achieving the desirable power. Moreover, by changing the relationship between β_1 , α_2 , and the effect size of population, the proposed method is shown to be robust for effect size larger than or equal to 1.5. Furthermore, the

simulation results give a relatively reliable guidance on how to decide the allocation ratio in order to achieve a desired statistical power for the trial. Hence, it is suggested that a simulation with the prespecified significance level, futility boundary and approximate sample size should be conducted before the trial in order to make a good selection of the allocation ratio of the sample.

2.7 Conclusion and Discussion

The proposed method for analysing two-stage seamless adaptive designs with both endpoint changes and population shifts relies on establishing relationships between endpoints and populations. These relationships allow prediction of the primary endpoint for interim analysis, after which data from both pre-shift and post-shift populations can be combined in the final analysis. This approach enables valid interpretation of treatment efficacy in the shifted population. As shown in Section 2.4, family-wise type I error can be effectively controlled, and simulation results in Section 2.6 demonstrate desirable power when the allocation ratio is appropriately chosen. In particular, the method exhibits good robustness when the effect size is at least 1.5, as the desired power can consistently be achieved through suitable sample allocation.

Beyond scenarios where population shifts result from protocol amendments, the proposed method is also applicable to trials in which shifts occur due to disease progression. For example, in metabolic dysfunction–associated steatohepatitis (MASH), the FDA (2018) recommends histological improvements such as steatohepatitis resolution or fibrosis reduction as surrogate markers predictive of longer-term clinical outcomes like survival or cirrhosis. Since these histological endpoints require invasive

procedures such as biopsy, recruitment may be challenging, making a two-stage seamless adaptive design attractive for MASH trials (Filozof et al., 2017). In such settings, changes in disease stage can be treated as a population shift, and the proposed framework can be applied to ensure valid final analysis and interpretation.

One potential limitation is that closed-form relationships between endpoints or populations may not always be available. In such cases, the proposed method would require modification or alternative modelling approaches. Nonetheless, when sufficiently accurate relationships can be specified, the framework provides a practical strategy to integrate information across stages and deliver reliable conclusions. While inference is focused on the shifted population, similar principles can be applied to evaluate treatment efficacy in the pre-shift population. Future work may further extend the method to settings with smaller effect sizes, broadening its applicability in complex adaptive trials.

Beyond these general methodological considerations, an important area of application lies in the development of biosimilar drug products. Biosimilar trials often require both pharmacokinetic (PK) and clinical evaluations, which traditionally proceed in a stepwise manner and can be time-consuming and resource intensive. To address these challenges, the next chapter introduces an innovative two-stage seamless PK-clinical adaptive design that integrates PK assessment with comparative clinical evaluation. This design builds on the principles established in Chapter 2 while tailoring them to the specific regulatory and scientific needs of biosimilar development, thereby

illustrating how methodological advances can translate into practical improvements in trial efficiency and decision-making.

3. A Proposal for the Use of Innovative Two-Stage PK-Clinical Adaptive Trial in Biosimilar Product Development

The methodological framework developed in Chapter 2 demonstrated how two-stage seamless adaptive designs can accommodate changes in both study endpoints and

patient populations. Under this framework, the present chapter (Chapter 3) turns to a concrete regulatory context where efficiency gains are especially critical: biosimilar drug development.

This chapter introduces an innovative two-stage seamless PK–clinical adaptive design that integrates pharmacokinetic assessment in stage 1 with comparative clinical evaluation in stage 2. By leveraging the biosimilarity index, the design enables early and reliable decisions on biosimilarity and directly responds to FDA’s call for more efficient development strategies. The integration of PK and clinical studies within a single adaptive framework has the potential to shorten development timelines, reduce costs, and accelerate regulatory approval of biosimilar products.

3.1 Introduction

After the expiration of patent protection of an innovator drug, pharmaceutical and/or biotech companies can make generic copies (e.g. generics for small-molecule products or biosimilars for large-molecule biological products) of it. The generic copies of the innovator drug product are less expensive than their brand-name counterparts, making them more affordable, and therefore benefit the target patient population.

For the approval of generic drugs, the United States Food and Drug Administration (US FDA) indicates that there should be an assessment of bioequivalence for demonstrating bioequivalence in terms of drug absorption and metabolism as compared to their reference products (FDA, 2003; FDA, 2021). In practice, bioequivalence assessment is typically conducted through the evaluation of pharmacokinetic (PK) responses such as C_{max} (maximum plasma concentration) and

AUC (area under the curve of plasma concentration versus time, representing total drug exposure integrated over time) (Chow & Liu, 2008). According to the Fundamental Bioequivalence Assumption, demonstrating bioequivalence in PK responses between a generic (test) drug product and its reference (innovator) product is indicative of the equivalence in their therapeutic effect.

In contrast to the generics, which typically contain the same active ingredient as the innovator drug product, biosimilars are made from living cells or microorganisms, which often comprise a mixture of numerous minor variants of a protein and are therefore more complex. Thus, during the production of biosimilars, it is not possible to make identical copies of the active ingredient of an innovator drug product in each dose or batch of the product. In 2015, the FDA recommended a stepwise approach for obtaining the totality of evidence to demonstrate biosimilarity between a proposed biosimilar (test) product and an innovator (reference) product (FDA, 2015). The stepwise approach includes but is not limited to the demonstration of analytical similarity for quality assurance/control, pharmacokinetic (PK) and/or pharmacodynamic (PD) similarity for drug absorption, metabolism and pharmacological activities, and clinical similarity for safety and efficacy, including the assessment of immunogenicity similarity to ensure immune safety.

FDA's recommended stepwise approach, however, has been criticized for being excessively time-consuming and not cost-effective, resulting in a delay in regulatory review and approval, and consequently, significant delays in bringing the affordable biosimilar product to the marketplace. Thus, the FDA (2022) calls for proposals to

improve the efficiency of the biosimilar development and review process. In response to a recent FDA's request for proposal (RFP), Mai and Chow proposes the use of an innovative two-stage seamless adaptive trial design that combines the assessment of PK similarity (the first stage) and clinical similarity (the second stage) to streamline the biosimilar development process.

The proposed two-stage seamless PK-clinical adaptive design (two-arm parallel design) addresses individual studies with the same objective (i.e., the demonstration of biosimilarity) but different endpoints (i.e., PK responses versus clinical endpoints). Biosimilarity is evaluated using the biosimilarity index (BI) as defined by Chow (2013), which operates under the Fundamental Biosimilarity Assumption that PK similarity predicts clinical similarity in comparative studies. This design allows for an early decision at the end of the first stage, based on PK data, regarding whether to proceed to the second stage. The principal aim of this design is to facilitate early and reliable decisions on biosimilarity at the conclusion of stage 1, thereby expediting the biosimilar development process. Should biosimilarity not be established in stage 1 and the decision is made to proceed to stage 2, adaptations such as sample size re-estimation based on the estimated margin may be introduced. Consequently, a notable advantage of this design is its potential to reduce trial duration, especially when biosimilarity is demonstrated in stage 1, thereby accelerating the overall development process, including the review and approval phases, of the biosimilar product.

In the next section, the test statistic based on reproducibility probability i.e., biosimilarity index (BI) will be derived. Section 3.3 will show the statistical method for

testing BI at stage 1 and stage 2 to make conclusion of biosimilarity. A simulation study will be conducted in Section 3.4 to evaluate the performance of the proposed method comparing to the stepwise approach. Concluding remarks are provided in the last section of this chapter.

3.2 Biosimilarity Index

For assessment biosimilarity between a proposed biosimilar (test) product and its innovator (reference) product, Chow (2013) proposed the use of biosimilarity index (BI), which is defined as reproducibility probability, under a pre-specified biosimilarity limits (θ_L, θ_U) .

Given that a specific alternative hypothesis H_1 is true, the power of a hypothesis test is the probability of correctly rejecting the null hypothesis H_0 , which is thereby a conditional probability. Unlike the power, which is evaluated given a specific alternative true value, the reproducibility probability (BI) is an unconditional probability calculated as the weighted average of the power across the range of all possible alternative true values.

The biosimilarity index can be derived based on the PK data collected in stage 1 under the Bayesian approach. The BI can be defined and expressed as the conditional probability of rejecting null hypothesis in the future trial given that the data observed is \mathbf{X} :

$$BI(\mathbf{X}) = P(\text{reject } H_0 | \mathbf{X}) = \int P(\text{reject } H_0 | \delta_{true}) P(\delta_{true} | \mathbf{X}) d\delta_{true}$$

3.2.1 Derivation of Biosimilarity Index and Statistical Properties

Now consider the biosimilarity test for PK parameter. Suppose n_T and n_R participants are recruited in stage 1 to test group T and reference group R respectively. Denote the data observed at stage 1 for PK parameter as $x_{i,j}$ where $i = T, R$ and $j = 1, \dots, n_i$. The distributions of log transformed data is assumed to be normally distributed: $\log x_{i,j} \sim N(\mu_i, \sigma_i^2)$.

The hypotheses of testing biosimilarity could be expressed as follows:

$$H_0: \mu_T - \mu_R \geq \ln \theta_L \text{ or } \mu_T - \mu_R \leq \ln \theta_U$$

$$H_1: \ln \theta_L < \mu_T - \mu_R < \ln \theta_U$$

Then a new test statistic will be derived by using the analysis method for the normal PK study with parallel design and model:

$$\log(\text{PK parameter}) = c + \beta \times \mathbb{1}_{\{\text{treatment}=\text{test drug}\}} + \epsilon_{\text{treatment}=\text{test drug}}.$$

In this article, without loss of generality and for simplicity, it is assumed that the random error term ϵ follows normal distribution, where the variances of the two treatment groups are assumed unequal.

Let β be the difference of means of $\log x_{T,j}$ and $\log x_{R,j}$, then the confidence interval would be $CI_{diff} = \beta \pm t(1 - \alpha, df_t) \times \sqrt{\sigma_R^2 + \sigma_T^2}$. Suppose α is the pre-specified confidence level, S_R^2, S_T^2 are the sample variances of the reference product and test product respectively. The pooled degrees of freedom df_t corresponding to pooled variance can thereby be calculated with the Welch-Satterthwaite equation.

$$df = \frac{(S_R^2/n_R + S_T^2/n_T)^2}{\frac{(S_R^2/n_R)^2}{n_R - 1} + \frac{(S_T^2/n_T)^2}{n_T - 1}}$$

The power function for testing a PK endpoint at significance level α is given

below:

$$\begin{aligned}
p_1(\mu_T, \mu_R) &= P \left(\ln \theta_L < \beta - t(1 - \alpha, df) \times \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}} \cap \beta + t(1 - \alpha, df) \times \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}} \right. \\
&\quad \left. < \ln \theta_U \right) \\
&= P \left(\ln \theta_L + t(1 - \alpha, df) \times \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}} < \beta < \ln \theta_U - t(1 - \alpha, df) \times \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}} \right) \\
&\approx \Phi \left(\frac{\ln \theta_U - (\mu_T - \mu_R)}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} - t(1 - \alpha, df) \right) \\
&\quad - \Phi \left(\frac{\ln \theta_L - (\mu_T - \mu_R)}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} + t(1 - \alpha, df) \right) \\
&= \Phi \left(-\frac{(\mu_T - \mu_R) - \ln \theta_U}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} - t(1 - \alpha, df) \right) \\
&\quad - \left(1 - \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_L}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} - t(1 - \alpha, df) \right) \right)
\end{aligned}$$

$$\begin{aligned}
&= \Phi \left(-\frac{(\mu_T - \mu_R) - \ln \theta_U}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} - t(1 - \alpha, df) \right) \\
&\quad + \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_L}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} - t(1 - \alpha, df) \right) - 1
\end{aligned}$$

In this article, for simplicity, only the unknown parameters μ_T and μ_R are assumed to be random with a known prior distribution $\pi(\mu_T, \mu_R)$, the variances are assumed to be fixed.

$$\begin{aligned}
BI(\mathbf{X}) &= P(\ln \theta_L < \mu_T - \mu_R < \ln \theta_U | \mathbf{X}) \\
&= \int \int p_1(\mu_T, \mu_R) \pi(\mu_T, \mu_R | \mathbf{X}) d\mu_T d\mu_R
\end{aligned}$$

If the variances are also random, then the $BI(\mathbf{X})$ has to be evaluated numerically (see e.g., Chow, 2013). When no available prior information can be obtained, different methods may be considered for the approximation of $BI(\mathbf{X})$, which includes but are not limited to the use of a non-informative prior, or the application of bootstrapping method (with replacement). With the use of non-informative prior, according to Chow (2013) (see also Shao & Chow, 2002),

$$\begin{aligned}
BI_1(\mathbf{X}) \approx & \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_U + t(1 - \alpha, df) \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} \right) \\
& + \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_L - t(1 - \alpha, df) \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} \right) - 1
\end{aligned}$$

Then the estimate of $BI_1(\mathbf{X})$ with a specific dataset \mathbf{X}_1 collected in stage 1 with a non-informative (uniform) prior $\pi(\mu_T, \mu_R) = 1$ is:

$$\begin{aligned}
\widehat{BI_1(\mathbf{X})} &= BI_1(\mathbf{X}_1) \\
&= \Phi \left(\frac{(\hat{\mu}_T - \hat{\mu}_R) - \ln \theta_U + t(1 - \alpha, df) \sqrt{\frac{S_R^2}{n_R} + \frac{S_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{S_R^2}{n_R} + \frac{S_T^2}{n_T}}} \right) \\
&+ \Phi \left(\frac{(\hat{\mu}_T - \hat{\mu}_R) - \ln \theta_L - t(1 - \alpha, df) \sqrt{\frac{S_R^2}{n_R} + \frac{S_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{S_R^2}{n_R} + \frac{S_T^2}{n_T}}} \right) - 1
\end{aligned}$$

The variance of $BI_1(\mathbf{X})$ should be derived for the hypothesis test afterwards. To derive $Var(BI_1(\mathbf{X}))$, first define following variables:

$$\begin{aligned}
Z_1 &= -\frac{\mu_T - \mu_R - \ln \theta_U}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}}, \\
Z_2 &= \frac{\mu_T - \mu_R - \ln \theta_L}{\sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}},
\end{aligned}$$

$$Z'_1 = \frac{Z_1 - t(1 - \alpha, df)}{\sqrt{2}},$$

$$Z'_2 = \frac{Z_2 - t(1 - \alpha, df)}{\sqrt{2}}.$$

Therefore,

$$BI_1(\mathbf{X}) = \Phi(Z'_1) + \Phi(Z'_2) - 1$$

$$Var(BI_1(\mathbf{X})) = Var[\Phi(Z'_1) + \Phi(Z'_2)]$$

By applying the Delta Method,

$$Var(BI_1(\mathbf{X})) = \phi^2(Z'_1)Var(Z'_1) + \phi^2(Z'_2)Var(Z'_2) + 2\phi(Z'_1)\phi(Z'_2)Cov(Z'_1, Z'_2)$$

Since Z_1 and Z_2 are standardised normal variables,

$$Var(Z'_1) = \frac{1}{2}Var(Z_1) = Var(Z'_2) = \frac{1}{2}Var(Z_2) = \frac{1}{2},$$

and

$$\begin{aligned} Cov(Z'_1, Z'_2) &= Cov\left(\frac{Z_1 - t(1 - \alpha, df)}{\sqrt{2}}, \frac{Z_2 - t(1 - \alpha, df)}{\sqrt{2}}\right) \\ &= \frac{1}{2}Cov(Z_1, Z_2) = -\frac{1}{2}. \end{aligned}$$

Finally,

$$\begin{aligned} Var(BI_1(\mathbf{X})) &= \frac{1}{2}\phi^2(Z'_1) + \frac{1}{2}\phi^2(Z'_2) - \phi(Z'_1)\phi(Z'_2) \\ &= \frac{1}{2}\phi^2\left(\frac{Z_1 - t(1 - \alpha, df)}{\sqrt{2}}\right) + \frac{1}{2}\phi^2\left(\frac{Z_2 - t(1 - \alpha, df)}{\sqrt{2}}\right) \\ &\quad - \phi\left(\frac{Z_1 - t(1 - \alpha, df)}{\sqrt{2}}\right)\phi\left(\frac{Z_2 - t(1 - \alpha, df)}{\sqrt{2}}\right) \end{aligned}$$

By plugging in the $\hat{\beta}$, S_R^2 , and S_T^2 value calculated from the collected dataset \mathbf{X}_1

into Z_1 and Z_2 , an estimation of variance $Var(\widehat{BI_1}(\mathbf{X})) = Var(BI_1(\mathbf{X}_1))$ can be obtained.

3.2.2 Remarks

The Biosimilarity Index (BI) has several advantages, making it a valuable tool in various scenarios. It is robust and applicable across different study endpoints, including pharmacokinetic (PK) responses, clinical endpoints, and biomarkers. Additionally, it allows different biosimilarity criteria in hypothesis testing and supports diverse study designs. The BI directly reflects the level of similarity, with a higher index indicating a greater degree of similarity, thereby meeting the FDA's requirement for products to be "highly similar."

Furthermore, Chow (2013) proposes that the boundary of an acceptable reproducibility probability p_0 should be determined through an R-R study, which evaluates the average biosimilarity of the same two reference products. The reproducibility probability for the R-R trial, based on the Average Bioequivalence (ABE) criterion, can be defined as

$$P_{RR} = P \left(\begin{array}{c} \textit{Concluding average biosimilarity of the same two reference} \\ \textit{products in a future trial given that the} \\ \textit{average biosimilarity is established in the first trial} \end{array} \right).$$

Since comparing the same two products typically yields a higher reproducibility probability, the criterion p_0 for an acceptable reproducibility probability when comparing a biosimilar to its innovator product can be chosen as a fraction of P_{RR} . For example, $p_0 = 0.8P_{RR}$. The selection of p_0 is considered a regulatory review issue. The higher value of p_0 indicates the higher degree of similarity. An appropriately selected p_0 may not only reflect the expected degree of similarity but also ensure that the early decision made in stage 1 is reliable.

3.3 Proposed Two-stage PK-Comparative Clinical Adaptive Trial

To apply the two-stage adaptive trial design (to a parallel trial), it is assumed that PK responses is predictive of the clinical endpoints. In this article, for illustration purposes, we will focus on continuous PK parameters such as C_{max} and AUC in PK study, and continuous clinical endpoint in the comparative clinical study.

If biosimilarity has been demonstrated in stage 1, we may stop the trial and conclude that the proposed biosimilar product is biosimilar to the reference product as there is no need to continue to stage 2 under the Fundamental Biosimilarity Assumption that PK responses are predictive of the clinical endpoints for assessment of safety and efficacy. Table 1 summarises the characteristics, corresponding test statistics and critical values at each stage.

Table 15: Characteristics of two-stage PK-Clinical proposal.

Note: BI_1 and BI_2 are biosimilarity index (reproducibility probability) at stage 1 and stage 2, respectively.

	Stage 1: PK	Stage 2: Clinical
Objective	Demonstration of Biosimilarity	Demonstration of Biosimilarity
Endpoints	PK response, e.g., C_{max} and AUC	Clinical endpoint
Population	Patients	Patients
Test Statistic	BI_1	BI_2
Corresponding p-value	p_1	p_2
α level	α_1	α_2

3.3.1 Analysis for Stage 1

Then to test for bioequivalence using the new statistic, following set of hypotheses will be used.

$$H_0: BI_1(\mathbf{X}) < p_0$$

$$H_1: BI_1(\mathbf{X}) \geq p_0$$

With the observed dataset \mathbf{X}_1 , the test statistic would be $\frac{BI_1(\mathbf{X}_1) - p_0}{\sqrt{\text{Var}(BI_1(\mathbf{X}_1))}}$ follows approximately normal distribution (when sample size larger than 30). Hence for the hypothesis testing, the null hypothesis H_0 would be rejected based on the p-value p_1 of this hypothesis test.

The decision on whether to continue to stage 2 is determined by following criteria:

Table 16: Decision boundaries at the end of stage 1.

Stop for biosimilar	$p_1 \leq \alpha_1$
Continue to stage 2	$\alpha_1 < p_1 < \beta_1$
Stop for not biosimilar	$p_1 \geq \beta_1$

3.3.2 Analysis for Stage 2

For the analysis of stage 2, there are two possible scenarios: 1) the exact predictive relationship between PK endpoint and the clinical endpoint exists but is unknown; 2) the exact predictive relationship between PK endpoint and the clinical endpoint is known. For the first scenario, the analysis of stage 2 would be based on the clinical endpoint data collected in stage 2; but for the second scenario, PK data collected in stage 1 can be used and combined with that in stage 2 for analysis. The method for

each scenario will be discussed in this section. Denote the data of primary clinical endpoint collected at stage 2 as $y_{i,k} \sim N(\mu_{y,i}, \sigma_{y,i}^2)$, with $k = 1, \dots, m_i$ and $i = T, R$. The sample variances of the clinical endpoint of reference product and test product are $S_{y,R}^2$ and $S_{y,T}^2$ respectively.

3.3.2.1 Endpoints Relationship is Not Well-established

The predictive relationship allows early decision making about biosimilarity in stage 1, and when the decision is to continue to stage 2, another hypothesis testing should be performed based on the clinical endpoint. Similar to stage 1, a test will be conducted on the Biosimilarity Index (BI) developed via following hypothesis:

$$H_0: \mu_{y,T} - \mu_{y,R} \geq \delta_L \text{ or } \mu_{y,T} - \mu_{y,R} \leq \delta_U$$

$$H_1: \delta_L < \mu_{y,T} - \mu_{y,R} < \delta_U$$

Following the same assumptions on the distribution on μ_T and μ_R , and the derivation in Section 2, the Biosimilarity Index for stage 2 when $\delta = \mu_{y,T} - \mu_{y,R}$ is:

$$BI_2(\mathbf{X}) \approx \Phi \left(\frac{\delta - \delta_U + t(1 - \alpha, df) \sqrt{\frac{\sigma_{y,R}^2}{m_R} + \frac{\sigma_{y,T}^2}{m_T}}}{\sqrt{2} \sqrt{\frac{\sigma_{y,R}^2}{m_R} + \frac{\sigma_{y,T}^2}{m_T}}} \right) + \Phi \left(\frac{\delta - \delta_L - t(1 - \alpha, df) \sqrt{\frac{\sigma_{y,R}^2}{m_R} + \frac{\sigma_{y,T}^2}{m_T}}}{\sqrt{2} \sqrt{\frac{\sigma_{y,R}^2}{m_R} + \frac{\sigma_{y,T}^2}{m_T}}} \right) - 1$$

Then conduct a hypothesis test on BI_2 :

$$H_0: BI_2(\mathbf{X}) < p_0$$

$$H_1: BI_2(\mathbf{X}) \geq p_0$$

With the dataset collected in stage 2 \mathbf{X}_2 , the estimated Biosimilarity Index is thereby $BI_2(\mathbf{X}_2)$, with $\mu_{y,T}$, $\mu_{y,R}$, $\sigma_{y,R}^2$ and $\sigma_{y,T}^2$ substituted by the estimations \bar{y}_T , \bar{y}_R , $S_{y,R}^2$ and $S_{y,T}^2$ generated from the sample data \mathbf{X}_2 , the corresponding p-value of this test is denoted as p_2 . To control the overall Type I error, p_2 should be compared with the critical value α_2 , see more detailed discussion on the calculation of α_2 in Section 2.4.

Table 17: Decision boundaries at the end of stage 2.

Stop for biosimilar	$p_2 \leq \alpha_2$
Stop for not biosimilar	$p_2 > \alpha_2$

3.3.2.2 Endpoints Relationship is Well-established

When there is a known predictive relationship between PK endpoint in stage 1 and clinical endpoint in stage 2, data collected in stage 1 can also be used to improve the estimation in stage 2 analysis. For simplicity of demonstration, assume there exist a known linear relationship between the PK endpoint and the clinical endpoint, and the random error term ϵ follows normal distribution (unequal variances are assumed for two treatment groups):

$$y_{i,j} = a + b \log x_{i,j} + \epsilon_{i,j}, \quad \mathbb{E}(\epsilon_{i,j} | \log x_{i,j}) = 0, \quad i \in \{T, R\}.$$

Superscript convention (stage labels). For any arm-level quantity, the superscripts indicate the stage of origin: $A_i^{(1)}$ comes from Stage 1 (PK/surrogate marker) and $A_i^{(2)}$ from Stage 2 (clinical/primary). In particular, $\overline{\log x_i^{(1)}}$ is the Stage-1 log-PK mean, $\bar{y}_i^{(2)}$ the Stage-2 clinical mean, $n_i^{(1)}$ and $m_i^{(2)}$ are the stage-specific arm sample sizes; O_i is the number of overlapping individuals observed in both stages ($0 \leq O_i \leq \min\{n_i^{(1)}, m_i^{(2)}\}$).

Patient-level variances (observed endpoints). We write

$$\sigma_{\log x,i}^2 := \text{Var}(\log x_{i,j}), \quad \sigma_{y,i}^2 := \text{Var}(y_{i,j}),$$

so that arm-mean variances are

$$\text{Var}(\overline{\log x}_i^{(1)}) = \frac{\sigma_{\log x,i}^2}{n_i^{(1)}}, \quad \text{Var}(\bar{y}_i^{(2)}) = \frac{\sigma_{y,i}^2}{m_i^{(2)}}.$$

Under the assumed linear relationship,

$$\sigma_{y,i}^2 = \text{Var}(y_{i,j}) = b^2 \sigma_{\log x,i}^2 + \sigma_{\epsilon,i}^2,$$

$$\text{Cov}(\log x_{i,j}, y_{i,j}) = b^2 \sigma_{\log x,i}^2.$$

PK-informed Stage-2 estimator.

Define the Stage-2 predictor from Stage-1 PK and the observed Stage-2 mean

$$A_i = a + b \overline{\log x}_i^{(1)}, \quad B_i = \bar{y}_i^{(2)},$$

and combine them as

$$\hat{\mu}_{y,i}^{GD} = \omega_i A_i + (1 - \omega_i) B_i, \quad 0 \leq \omega_i \leq 1.$$

Both A_i and B_i are unbiased for $\mu_{y,i} = \mathbb{E}(y_{i,j})$; therefore $\hat{\mu}_{y,i}^{GD}$ is unbiased

(Appendix C).

Variance and optimal weight.

The within-arm cross-stage covariance of means depends on overlap:

$$\text{Cov}(\overline{\log x}_i^{(1)}, \bar{y}_i^{(2)}) = \begin{cases} \frac{b \sigma_{\log x,i}^2}{n_i^{(1)}} & \text{fully paired } (O_i = n_i^{(1)} = m_i^{(2)}), \\ \frac{b \sigma_{\log x,i}^2 O_i}{n_i^{(1)} m_i^{(2)}} & \text{partial overlap } (0 \leq O_i \leq \min\{n_i^{(1)}, m_i^{(2)}\}), \\ 0 & \text{no overlap } (O_i = 0). \end{cases}$$

Hence

$$\text{Var}(A_i) = \frac{b^2 \sigma_{\log x,i}^2}{n_i^{(1)}}, \quad \text{Var}(B_i) = \frac{\sigma_{y,i}^2}{m_i^{(2)}}, \quad \text{Cov}(A_i, B_i) = b \text{Cov}(\overline{\log x}_i^{(1)}, \bar{y}_i^{(2)}),$$

and the variance and optimal weight are

$$\text{Var}(\hat{\mu}_{y,i}^{GD}) = \omega_i^2 \text{Var}(A_i) + (1 - \omega_i)^2 \text{Var}(B_i) + 2\omega_i(1 - \omega_i) \text{Cov}(A_i, B_i),$$

$$\omega_i^* = \frac{\text{Var}(B_i) - \text{Cov}(A_i, B_i)}{\text{Var}(A_i) + \text{Var}(B_i) - 2\text{Cov}(A_i, B_i)}.$$

Useful specialisations.

- Fully paired ($O_i = n_i^{(1)} = m_i^{(2)}$):

$$\text{Var}(\hat{\mu}_{y,i}^{GD}) = \frac{1}{n_i^{(1)}} [b^2 \sigma_{\log x,i}^2 + (1 - \omega_i)^2 \sigma_{\epsilon,i}^2].$$

- Partial Overlap:

$$\text{Var}(\hat{\mu}_{y,i}^{GD}) = b^2 \sigma_{\log x,i}^2 \left(\frac{\omega_i^2}{n_i^{(1)}} + \frac{2\omega_i(1 - \omega_i)O_i}{n_i^{(1)}m_i^{(2)}} \right) + (1 - \omega_i^2) \frac{\sigma_{y,i}^2}{m_i^{(2)}}.$$

- No overlap ($O_i = 0$):

$$\text{Var}(\hat{\mu}_{y,i}^{GD}) = b^2 \sigma_{\log x,i}^2 \frac{\omega_i^2}{n_i^{(1)}} + (1 - \omega_i^2) \frac{\sigma_{y,i}^2}{m_i^{(2)}}.$$

If the cross-stage dependence vanishes $\text{Cov}(A_i, B_i) = 0$; e.g., $O_i = 0$), the optimal

weight collapses to the Graybill-Deal (GD) weight

$$\hat{\omega}_i = \frac{\frac{n_i^{(1)}}{b^2 S_{\log x,i}^2}}{\frac{n_i^{(1)}}{b^2 S_{\log x,i}^2} + \frac{m_i^{(2)}}{S_{y,i}^2}}.$$

From arm estimators to the treatment contrast.

With independent arms,

$$\hat{\Delta}^{GD} = \hat{\mu}_{y,T}^{GD} - \hat{\mu}_{y,R}^{GD}, \quad SE(\hat{\Delta}^{GD}) = \sqrt{\widehat{\text{Var}}(\hat{\mu}_{y,T}^{GD}) + \widehat{\text{Var}}(\hat{\mu}_{y,R}^{GD})}.$$

The rest of the hypothesis testing steps in this scenario will be the same as that in

Section 3.3.2.1, the only difference in calculation is to substitute $\sqrt{\frac{S_{y,R}^2}{m_R} + \frac{S_{y,T}^2}{m_T}}$ by

$\sqrt{\widehat{Var}(\hat{\mu}_{y,T}^{GD}) + \widehat{Var}(\hat{\mu}_{y,R}^{GD})}$, and to substitute \bar{y}_T, \bar{y}_R by $\hat{\mu}_{y,T}^{GD}$ and $\hat{\mu}_{y,R}^{GD}$.

3.4 Type I Error Control

3.4.1 Setup and Notation

Let p_1 and p_2 be the one-sided p-values from the BI tests at Stage 1 (PK) and Stage 2 (clinical). Define the corresponding Z-scores

$$T_k = \Phi^{-1}(1 - p_k), \quad k \in \{1, 2\},$$

so that larger T_k favours biosimilarity. Stage 1 proceeds to Stage 2 only if

$$\mathcal{P} = \{z_{F_1} < T_1 < z_{E_1}\}, \quad z_{E_1} = \Phi^{-1}(1 - \alpha_1), \quad z_{F_1} = \Phi^{-1}(1 - \beta_1).$$

Let the familywise error target be α and define the remaining after Stage 1 as

$$\alpha_{rem} = \alpha - \Pr_{H_0}(p_1 \leq \alpha_1).$$

From exposition, when p_1 is (approximately) uniform under H_0 , $\Pr_{H_0}(p_1 \leq \alpha_1) \approx \alpha_1$

and $\Pr_{H_0}(\mathcal{P}) \approx \beta_1 - \alpha_1$.

Conditional-Error Rule

Under H_0 , the joint behaviour of (T_1, T_2) is well approximated by a bivariate normal model

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad -1 \leq \rho \leq 1,$$

where ρ quantifies the stage-to-stage correlation (e.g., due to data reuse or shared nuisance estimation). Conditioning on the observed interim value $T_1 = t$ and on proceeding, the Stage-2 distribution is

$$T_2|T_1 = t \sim \mathcal{N}(\rho t, 1 - \rho^2).$$

Spending the pre-proceed budget

$$q = \frac{\alpha_{rem}}{\Pr(\mathcal{P})_{H_0}} \approx \frac{\alpha - \alpha_1}{\beta_1 - \alpha_1}.$$

via the moving boundary

$$c_2(T_1) = \rho T_1 + \sqrt{1 - \rho^2} z_{1-q} \Leftrightarrow p_2 \leq 1 - \Phi\left(\rho T_1 + \sqrt{1 - \rho^2} z_{1-q}\right)$$

(reject at Stage 2 if $T_2 \geq c_2(T_1)$) ensures

$$\Pr_{H_0}(\text{Stage 2 reject}, \mathcal{P}) = q \Pr_{H_0}(\mathcal{P}) = \alpha_{rem}$$

and hence overall type I error α . When $\rho = 0$, $c_2(T_1) = z_{1-q}$ (constant threshold).

3.4.2 Estimating ρ in Practice

3.4.2.1 Design-Based plug-in (when Stage 2 reuses Stage 1)

With

$$\begin{aligned} \hat{\mu}_{y,i}^{GD} &= \omega_i(a + b \overline{\log x_i^{(1)}}) + (1 - \omega_i)\bar{y}_i^{(2)}, \\ \Delta_1 &= \overline{\log x_T^{(1)}} - \overline{\log x_R^{(1)}}, \quad \Delta_2^{GD} = \hat{\mu}_{y,T}^{GD} - \hat{\mu}_{y,R}^{GD}, \end{aligned}$$

use

$$\hat{\rho} \approx \frac{\text{Cov}(\Delta_1, \Delta_2^{GD})}{\sqrt{\text{Var}(\Delta_1)\text{Var}(\Delta_2^{GD})}} \quad (\text{under } H_0),$$

with $\text{Var}(\hat{\mu}_{y,i}^{GD})$ computed from the covariance-aware variance above and $\text{Cov}(\overline{\log x_i^{(1)}}, \bar{y}_i^{(2)})$ chosen by overlap (fully overlap/ partial/ independent).

3.4.2.2 Pilot-based Monte Carlo

Simulate the planned analysis under H_0 (target $n^{(1)}, m^{(2)}, O, \omega$; variances; exact BI computations and decision rules). For each replicate compute

$$T_1 = \Phi^{-1}(1 - p_1), \quad T_2 = \Phi^{-1}(1 - p_2)$$

then set $\hat{\rho} = Cor(T_1, T_2)$ and use this in $c_2(T_1)$.

3.5 Simulation

In this section, numerical simulation techniques are employed to assess the performance of the traditional confidence interval approach, which is commonly used in bioequivalence studies, in comparison to the proposed biosimilarity index (*BI*) approach under varying sample size and variance conditions.

The main purpose of the proposed design is to enable early and reliable decisions on biosimilarity at the end of stage 1, thereby expediting the biosimilar development process. Accordingly, the main objective of this simulation is to compare the traditional confidence interval approach with the *BI* approach, providing insights into the selection of a reliable and efficient similarity threshold p_0 for critical decision-making at the conclusion of stage 1. The simulation results could provide guidance for the selection of an acceptable boundary p_0 for assessing biosimilarity between the two products.

All simulations in this section are conducted under the following assumptions:

(i) the means (after log-transformation) of two treatment groups (1:1 randomisation ratio) are equal, (ii) the overall significance level $\alpha = 0.05$, and (iii) stage 1 significance level $\alpha_1 = 0.025$.

3.5.1 Corresponding p_0^* at Different Sample Sizes when Using CI Approach

Firstly, we aim to investigate the value of p_0^* when biosimilarity is concluded using the *CI* approach. Here, p_0^* is defined as the mean of the biosimilarity indices (*BI*) corresponding to the instances where the *CI* approach concludes biosimilarity. Specifically, $p_0^* = \text{mean}(BIs)$, where *BIs* represents the list of *BI* values associated with the *CI*-derived biosimilarity conclusions. For illustration purpose, it is assumed that $\sigma = 0.3$ and the p_0^* for each sample size is computed using 5000 simulations.

Table 18: Corresponding p_0^* of *CI* approach.

N	σ	p_0^*
40	0.3	0.0905
50	0.3	0.0925
60	0.3	0.1131
70	0.3	0.1206
80	0.3	0.1559
100	0.3	0.2398
150	0.3	0.4357
200	0.3	0.5737
300	0.3	0.7439

Table 18 illustrates that the corresponding p_0^* for concluding biosimilarity using the confidence interval (*CI*) approach is very small. The results also indicate that achieving a high reproducibility probability (p_0^*) with the *CI* approach necessitates a sufficiently large sample size. This requirement highlights a critical consideration (trade-off) in the application of the *CI* approach: smaller sample sizes may lead to an increased

risk of false positives. Specifically, when the sample size is inadequate, the *CI* approach may overestimate the likelihood of biosimilarity, thereby elevating the probability of incorrectly concluding that two products are biosimilar. As n_T and n_R decrease, the standard error increases, causing the confidence interval (*CI*) to widen. This, in turn, raises the likelihood that the *CI* will overlap with the biosimilarity margin, even if the true difference $\mu_T - \mu_R$ is outside the margin, ultimately leading to an incorrect conclusion of biosimilarity and increasing the risk of false positives.

This insight underscores the importance of robust study design and sample size determination in achieving accurate and reproducible outcomes in biosimilarity assessments.

3.5.2 Comparing Conclusions on Biosimilarity with Different Approaches

With the same sample size, variance and boundary p_0 , the conclusions on biosimilarity using *BI* approach and confidence interval (*CI*) approach when p_0 varies are compared in Table 19.

Table 19: Biosimilarity concluded from *CI* and *BI* approach with varying p_0 .

μ_R	μ_T	σ	N	p_0 = P_{RR}	<i>BI</i>	<i>p value</i> (<i>BI</i>)	<i>BI</i> result s	<i>CI</i>	<i>CI</i> results
1	1	0.3	300	0.4	0.7928	1.024e-04	Yes	[0.94, 1.15]	Yes
1	1	0.3	300	0.5	0.7928	2.823e-03	Yes	[0.94, 1.15]	Yes

1	1	0.3	300	0.6	0.7928	3.420e-02	No	[0.94, 1.15]	Yes
1	1	0.3	300	0.7	0.7928	1.902e-01	No	[0.94, 1.15]	Yes
1	1	0.3	300	0.8	0.7928	5.272e-01	No	[0.94, 1.15]	Yes

When the sample size and variance are fixed, the biosimilarity index (BI) and confidence interval (CI) will remain constant, leading to an unchanging conclusion on biosimilarity based on the CI approach. However, varying p_0 results in changes to the p-value of the BI approach, thereby altering the conclusion on biosimilarity according to the p-value. It is evident that if a higher p_0 (i.e., higher reproducibility probability) is expected, the BI approach will conclude that the two drugs are not biosimilar, whereas the CI approach consistently concludes biosimilarity. Thus, it is reasonable to assert that the BI approach is more stringent than the CI approach in determining biosimilarity.

3.5.3 Varying Sample Sizes

Simulation in this section is performed with $p_0 = 0.4$, $\sigma = 0.3$, and varying sample sizes. Table 20 summarises the results of *BI*, the corresponding p-values, the conclusion on biosimilarity using *BI* approach and confidence interval (*CI*) approach.

Table 20: Simulation results with varying sample sizes.

μ_R	μ_{trt}	σ	N	p_0 = P_{RR}	<i>BI</i>	<i>p value</i> (<i>BI</i>)	<i>BI</i> result s	<i>CI</i>	<i>CI</i> result s
---------	-------------	----------	-----	---------------------	-----------	---------------------------------	--------------------------	-----------	--------------------------

1	1	0.3	30	0.5	0	1	No	[0.77, 1.49]	No
1	1	0.3	60	0.5	0.0270	8.197e-01	No	[0.99, 1.51]	No
1	1	0.3	80	0.5	0.1655	9.285e-01	No	[0.89, 1.29]	No
1	1	0.3	100	0.5	0.2808	9.148e-01	No	[0.81, 1.13]	Yes
1	1	0.3	150	0.5	0.5258	3.988e-01	No	[0.90, 1.18]	Yes
1	1	0.3	200	0.5	0.6442	1.238e-01	No	[0.92, 1.17]	Yes
1	1	0.3	300	0.5	0.7928	2.822e-03	Yes	[0.94, 1.15]	Yes
1	1	0.3	350	0.5	0.8409	8.560e-05	Yes	[0.95, 1.15]	Yes
1	1	0.3	400	0.5	0.9189	3.015e-35	Yes	[0.94, 1.12]	Yes

Table 20 demonstrates that, given the same variance, the biosimilarity index (*BI*)

approach necessitates a larger sample size compared to the traditional confidence interval (*CI*) approach to conclude biosimilarity. This finding underscores the importance of ensuring an adequate sample size in stage 1 of the study to achieve a reliable conclusion of biosimilarity when employing the *BI* approach. The requirement for a larger sample size arises from the more stringent nature of the *BI* approach, which aims to enhance the robustness and reproducibility of the biosimilarity assessment. Consequently, researchers must carefully consider the sample size during the study design phase to meet the necessary criteria for concluding biosimilarity using the *BI* method.

3.5.4 Determining the Appropriate Sample Size for BI Approach Given p_0

The simulations conducted in this section aim to offer guidance on selecting an appropriate sample size for stage 1 when employing the Biosimilarity Index (*BI*) approach, tailored to each value of p_0 and variance σ^2 , under the condition that $\mu_T = \mu_R$ (i.e., the means of the test and reference groups are equal). The objective is to determine the sample size needed to achieve statistically significant results with the *BI* approach under various conditions. For illustrative purposes, we consider a specific scenario where the standard deviation σ is set to 0.3. By exploring this example, we seek to demonstrate how different values of p_0 influence the required sample size, thereby providing practical insights into the design of studies using the *BI* approach.

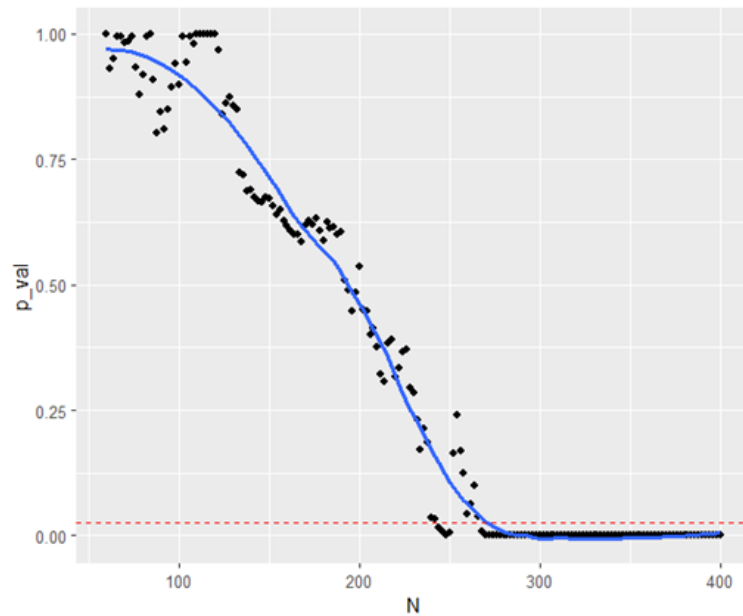


Figure 8: Simulation results when $p_0 = 0.4$.

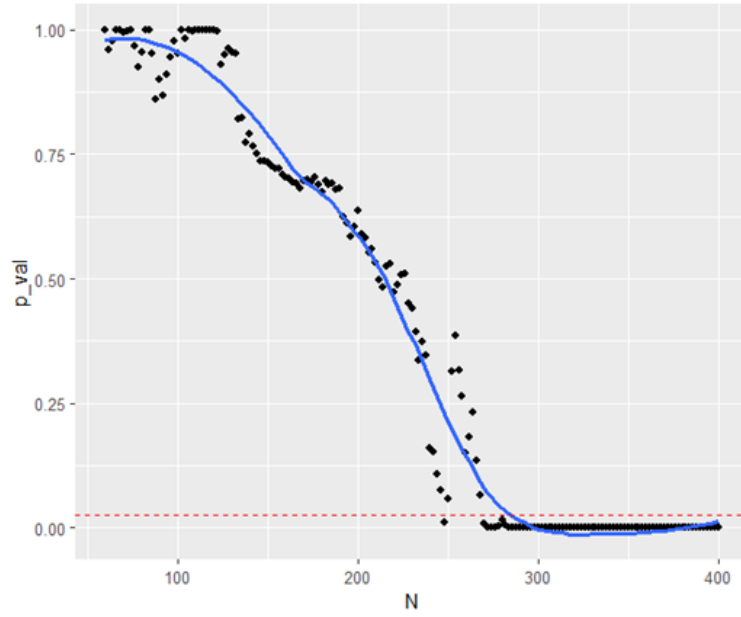


Figure 9: Simulation results when $p_0 = 0.5$

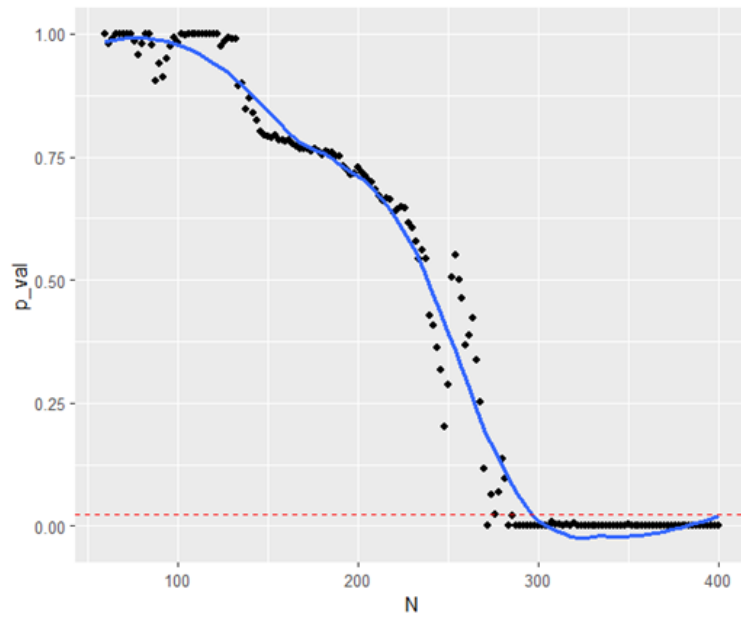


Figure 10: Simulation results when $p_0 = 0.6$.

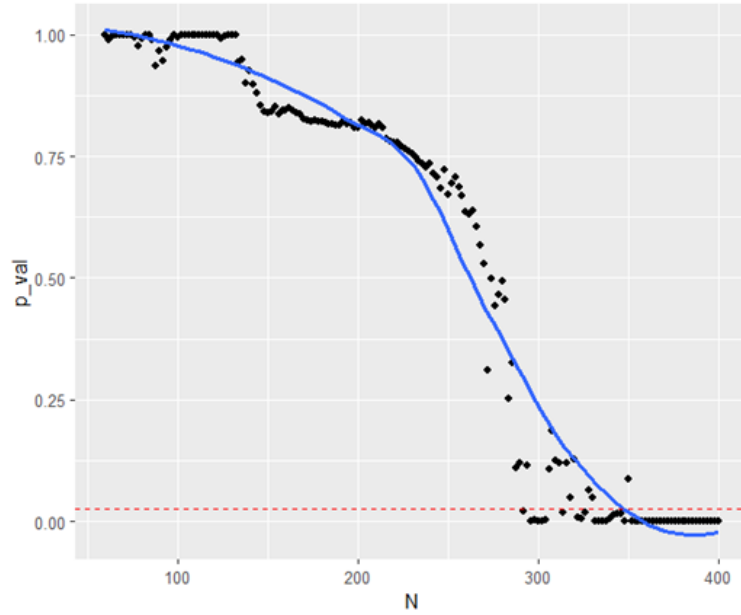


Figure 11: Simulation results when $p_0 = 0.7$.

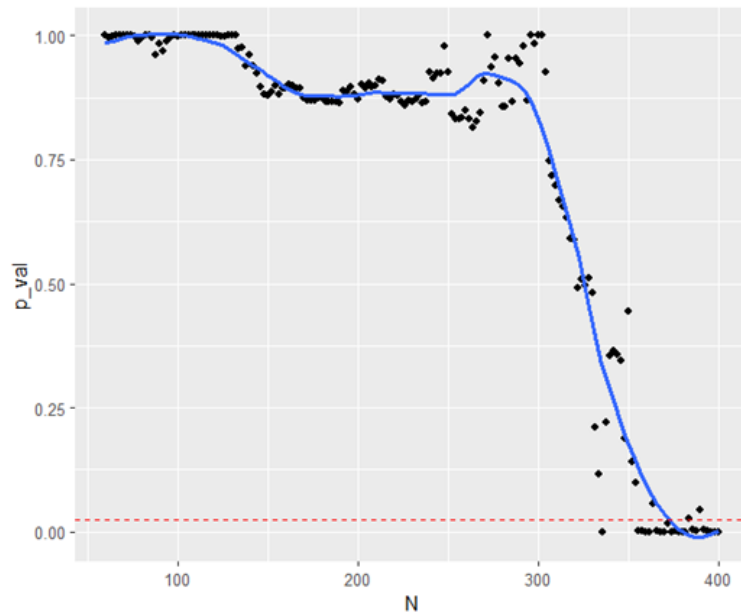


Figure 12: Simulation results when $p_0 = 0.8$.

Figures 8 through 12 illustrate the p-values obtained from the Biosimilarity Index (*BI*) approach for various values of p_0 across different sample sizes, with the red dotted line representing the significance level $\alpha_1 = 0.025$. In these figures, the sample size that

ensures nearly all p-values fall below the red dotted line indicates the required sample size for a specific p_0 , given that $\sigma = 0.3$.

From the trends observed in these figures and based on theoretical intuition, it is evident that higher values of the reproducibility probability p_0 necessitate larger sample sizes. This relationship underscores the need for more substantial sample sizes to achieve a stringent assessment of biosimilarity as the expected reproducibility probability increases.

To address variations in different scenarios, additional simulations can be conducted with varying values of σ^2 . These simulations would produce a new set of figures, offering further guidance on selecting the appropriate sample size under different variance conditions. This approach ensures that the sample size recommendations are adaptable to a range of statistical and practical contexts in biosimilarity assessments.

3.5.5 Determining the Appropriate p_0 for *BI* Approach Given Sample Size

The simulations presented in this section aim to provide guidance on selecting an appropriate p_0 for the Biosimilarity Index (BI) approach, given a specific sample size (for stage 1) and variance (σ^2), under the condition that $\mu_T = \mu_R$ (i.e., the means of the test and reference groups are equal). The objective is to identify the p_0 value that best balances the reproducibility probability and the probability of success of the trial using the BI approach. Here, the probability of success is defined as the probability of concluding biosimilarity based on the BI approach.

For illustrative purposes, we consider a scenario where the standard deviation (σ) is set to 0.3, and the probability of success for each p_0 is computed using 5000 simulations. By examining this example, we aim to demonstrate how different values of p_0 and sample sizes influence the probability of success of the trial. This analysis provides practical insights into the design of studies utilizing the BI approach.

Table 21: Probability of success of the trial with varying p_0 and sample sizes.

Sample size	p_0	Probability of success
150	0.3	0.1762
	0.35	0.1242
	0.4	0.0700
	0.45	0.0126
	0.5	0.0020
	0.55	0.0004
	0.6	0.0002
	0.65	0
	0.7	0
	0.75	0
	0.8	0
	0.85	0
200	0.3	0.4000
	0.35	0.3468
	0.4	0.2896
	0.45	0.1700

	0.5	0.1072
	0.55	0.0476
	0.6	0.0142
	0.65	0.0034
	0.7	0.0002
	0.75	0
	0.8	0
	0.85	0
	0.9	0
250	0.3	0.5970
	0.35	0.5492
	0.4	0.4970
	0.45	0.4386
	0.5	0.3760
	0.55	0.3062
	0.6	0.2392
	0.65	0.1568
	0.7	0.0748
	0.75	0.0202
	0.8	0.0026
	0.85	0
	0.9	0
300	0.3	0.7456
	0.35	0.7056
	0.4	0.6618

	0.45	0.6162
	0.5	0.5608
	0.55	0.4976
	0.6	0.4238
	0.65	0.3478
	0.7	0.2608
	0.75	0.1590
	0.8	0.0628
	0.85	0.0068
	0.9	0
350	0.3	0.8472
	0.35	0.8198
	0.4	0.7818
	0.45	0.7480
	0.5	0.7064
	0.55	0.6570
	0.6	0.5924
	0.65	0.5222
	0.7	0.4338
	0.75	0.3372
	0.8	0.2174
	0.85	0.0828
	0.9	0.0052

Table 21 depicts the probability of successfully concluding biosimilarity using the BI approach across varying sample sizes and p_0 values. This table provides a

reference for selecting a p_0 value that reflects the anticipated level of biosimilarity (reproducibility probability of biosimilarity results) while maintaining an acceptable probability of passing the biosimilarity test.

The trend observed in this table indicates that setting a higher reproducibility probability boundary (p_0) increases the difficulty for the new drug to pass the biosimilarity test, thereby reducing the probability of success in establishing biosimilarity between the biosimilar and the reference (innovator) drug. Additionally, it is evident that a higher p_0 necessitates a larger sample size to achieve a higher probability of success. This method for determining p_0 thus ensures reliable biosimilarity conclusions while avoiding excessively stringent criteria, ultimately facilitating the robust evaluation and approval of biosimilar drugs.

3.6 Discussion and Concluding Remarks

In clinical development, the two-stage seamless adaptive trial design, which amalgamates two trials that were previously designed and conducted independently into a single trial, has garnered substantial scholarly interest. A typical instance is the phase 2/3 seamless adaptive clinical trial. This design may encompass distinct study objectives, endpoints (potentially of varying data types), and target patient populations at different stages (Mai and Chow, 2024). In line with this methodology, and under the Fundamental Biosimilarity Assumption that pharmacokinetic (PK) similarity is indicative of clinical similarity, we propose a two-stage seamless trial design that integrates a PK study (for evaluating drug absorption) with a comparative clinical study (for assessing safety and efficacy) within the context of biosimilar drug development.

The aim of this proposed method is to streamline the review and approval process of the biosimilar development.

The utilisation of a biosimilarity index provides an effective solution to the challenge of quantifying "highly similar" relationships, as it captures the level of similarity between test products. The biosimilarity index is robust and exhibits insensitivity to the specific similarity margins used in biosimilar drug development. This robustness extends across various domains of biosimilarity assessments, including analytical studies, pharmacokinetic (PK) evaluations, and clinical trials. By offering a consistent measure of similarity irrespective of the margin thresholds, the biosimilarity index ensures a more reliable and comprehensive evaluation of biosimilar products, thereby enhancing the overall assessment process in biosimilar drug development.

One of the primary objectives of our proposed design is to facilitate the reliable early determination of biosimilarity or not. This objective is accomplished by establishing a framework that enables accurate and trustworthy conclusions about biosimilarity at an earlier stage of the development process. By implementing robust statistical methods and stringent criteria within the design, we aim to provide a solid foundation for making informed decisions regarding the biosimilarity of products before the completion of all study stages. This early decision-making capability is critical for streamlining the biosimilar development process, potentially reducing time and costs while ensuring that only products meeting stringent biosimilarity standards progress to subsequent phases.

The findings in this paper underscore that the Biosimilarity Index (BI) approach exhibits greater stringency compared to the confidence interval (CI) approach when concluding biosimilarity between two products. This increased rigor provides a higher level of assurance in the reproducibility of biosimilarity results in subsequent trials. The enhanced stringency of the BI approach establishes a more rigorous threshold for concluding biosimilarity, thereby reducing the likelihood of false positives. Consequently, this method ensures that only truly biosimilar products are deemed equivalent, which is crucial for maintaining the integrity and reliability of biosimilarity assessments.

The simulation provides guidance for selecting the appropriate sample size for the Biosimilarity Index (BI) approach. By evaluating various sample sizes and their impact on BI p-values, the simulation identifies the minimum sample size needed to achieve statistically significant results. This ensures that the sample size meets the BI approach's stringent criteria, thereby enhancing the reliability of biosimilarity conclusions. These insights inform study design, ensuring efficient resource allocation and adequate study power to detect true biosimilarity.

The simulation in Section 3.4 also offers essential insights for determining the optimal p_0 for a given sample size, ensuring a balance between attaining the expected level of reproducibility in biosimilarity results and maintaining an acceptable probability of success. This balance is essential in ensuring that the biosimilarity conclusions drawn from the trial are reliable and statistically robust, while avoiding overly stringent criteria that could unduly diminish the trial's probability of success.

This approach to selecting p_0 helps in designing more efficient and effective biosimilarity studies, ultimately contributing to the robust evaluation and approval of biosimilar drugs.

This study operates under the assumption that participants in both stages of the study are sourced from the same population. However, if the study involves different populations in the two stages, further adjustments will be required if the trial advances to the second stage. These adjustments are necessary to account for potential differences between the populations, which may impact the validity and reliability of the results. More details on implementing such modifications may be found in Mai and Chow (2024) as a reference.

Moreover, further investigations may explore whether using an unequal randomization ratio could reduce the total sample size required. By examining and optimizing different allocation ratios between treatment and control groups, it may be possible to design more efficient studies that maintain statistical power and reliability while requiring fewer participants. This research could significantly enhance the cost-effectiveness and feasibility of biosimilarity assessments.

Further expansion on the biosimilarity index is also necessary to facilitate its correct usage when evaluating biosimilarity and, in particular, in determining the appropriate similarity threshold for such testing. This motivates the methodological developments in the next chapter, which focus on establishing a rigorous statistical framework based on the Relative Biosimilarity Index (RBI) for defining thresholds and decision rules in biosimilar evaluation.

4. An Innovative Method Based on Relative Biosimilarity Index for Assessing Biosimilar Drug Products

The two-stage seamless PK–clinical adaptive design presented in Chapter 3 demonstrated how adaptive methods can streamline biosimilar development by integrating pharmacokinetic and clinical assessments within a unified trial. That design highlighted the utility of the biosimilarity index (BI) as a robust metric for guiding early decision-making. However, it also underscored an important methodological challenge: the need for principled guidance on selecting appropriate decision boundaries in biosimilarity testing. In particular, the choice of the margin in the hypothesis $H_0: BI(X) < p_0$ directly determines the stringency of the evaluation, influencing both the probability of declaring biosimilarity and the credibility of the resulting conclusion.

This chapter addresses that gap by extending the biosimilarity index (BI) framework through the development of the Relative Biosimilarity Index (RBI). The RBI, defined as the ratio of the BI to the reference-to-reference similarity, explicitly accounts for variability inherent in the reference product and thereby provides a more stable and informative basis for inference. For regulatory review and approval, this distinction matters: BI implies product specific decision thresholds that move with the variability of the reference, whereas RBI normalises for that variability and thus allows a single,

standardised similarity threshold to be prespecified and applied consistently across products, studies and therapeutic areas. In practice, RBI therefore facilitates more consistent, transparent and predictable determinations of biosimilarity at the time of regulatory submission. Building on this foundation, a systematic methodology is proposed for determining the corresponding similarity threshold δ , drawing on established equivalence margins and incorporating variability adjustments to ensure both statistical rigor and practical interpretability. Multiple inference procedures, including Fieller's method, logit transformation, beta regression, and probability-based approaches, are evaluated within this unified decision framework. Taken together, these developments establish an extended methodology that offers structured guidance on similarity threshold selection, bridging theoretical considerations with practical implementation in biosimilar evaluation.

4.1 Introduction

In recent years, the development of biosimilar drug products has received much attention. The purpose is to establish a framework that enhances the biosimilar review and approval process, thereby potentially enhancing the affordability and expediting the availability of biosimilar products, while ensuring the safety, efficacy, and quality of the approved biosimilar products. The framework is to demonstrate biosimilarity between the proposed biosimilar (test) product and the innovative (reference) product, which involves comparing the two products across analytical critical quality attribute (CQA), pharmacokinetic and pharmacodynamic (PK/PD), and clinical endpoints (FDA, 2015).

However, this process can be time-consuming and resource-intensive, creating a significant bottleneck for the timely approval of biosimilars.

To streamline biosimilarity evaluation and accelerate decision-making, Chapter 4 proposed a two-stage seamless adaptive design that integrates the PK and clinical studies into a single trial under the Fundamental Biosimilarity Assumption which assumes that the PK/PD similarity is predictive of clinical similarity. This innovative design enables early determination of biosimilarity at the end of stage 1, offering the potential to reach definitive conclusions on biosimilarity at an earlier phase of development. By eliminating the need for separate studies, the two-stage seamless adaptive design reduces redundancy, optimises resource utilization, and enhances the overall efficiency of the biosimilar development.

Chapter 4 discussed the use of biosimilarity index (*BI*), a novel metric proposed to provide a more flexible and comprehensive framework for assessment of biosimilarity. The biosimilarity index is robust across various domains of biosimilarity assessments and enables more straightforward interpretation of biosimilarity compared to conventional bioequivalence testing approaches.

As indicated by Chapter 4, biosimilarity evaluation lies the challenge of establishing appropriate decision-making thresholds, specifically the selection of the margin in the hypothesis $H_0: BI(\mathbf{X}) < p_0$, which determines the level of similarity required between the biosimilar and its reference product. The choice of p_0 directly influences the stringency of the test and the probability of passing, thereby affecting both clinical decisions and regulatory outcomes. This challenge highlights the need for a

rigorous, methodological approach to defining this boundary, specifically tailored to the context of the biosimilarity index. Building upon the biosimilarity index framework, this paper addresses the issue of similarity threshold selection. We propose a structured approach for testing biosimilarity, determining decision boundaries, and refining theoretical considerations for innovative boundary setting by drawing on traditional boundary-setting strategies. As a result, in this article, we propose a new metric, i.e., the relative biosimilarity index, which is defined as $RBI = BI/P_{RR}$ to account for variability inherent in the reference product. In addition, we propose to evaluate biosimilarity by testing the null hypothesis that $H_0: RBI \leq \delta$, and propose a systematic methodology for determining the corresponding decision threshold (δ).

Thus, the purpose of this study is threefold. First, we propose a new metric, the Relative Biosimilarity Index (RBI), which extends the biosimilarity index framework by normalizing against the variability inherent in the reference product. Second, we discuss and compare multiple statistical testing procedures for evaluating RBI, including Fieller's method, logit transformation, beta regression, and probability-based approaches, all implemented under a common decision framework. Third, we introduce a structured methodology for selecting the decision threshold δ , grounded in regulatory equivalence margins and adjusted for between-subject variability. A comprehensive simulation study is conducted to assess how the choice of δ , sample size, and testing procedure impacts type I error control and statistical power. The overarching goal is to bridge the gap between theoretical innovation and practical implementation in

biosimilar drug development, ultimately enhancing the rigor and efficiency of biosimilar drug development.

4.2 Relative Biosimilarity Index (RBI)

To assess biosimilarity between a proposed biosimilar product and its innovative (reference) product, we propose the following relative biosimilarity index (RBI):

$$RBI = \frac{BI(\mathbf{X})}{P_{RR}}$$

where $BI(\mathbf{X})$ is the reproducibility probability, which is defined as biosimilarity index derived from the observed data \mathbf{X} . P_{RR} is the BI obtained with data \mathbf{X}_{ref} , which the BI comparing the reference group with itself, and RBI can thereby be obtained. In what follows, for a given biosimilar study, the derivation of BI will be briefly described.

4.2.1 Biosimilarity Index

As discussed in Chapter 3, to assess biosimilarity between a proposed biosimilar (test) product and its innovator (reference) product, Chow (2013) introduced the biosimilarity index (BI), defined as the reproducibility probability within pre-specified biosimilarity limits (θ_L, θ_U) . The BI can be expressed as the conditional probability of rejecting null hypothesis in the future trial given that the data observed is \mathbf{X} :

$$BI(\mathbf{X}) = P(\text{reject } H_0 | \mathbf{X}) = \int P(\text{reject } H_0 | \delta_{true}) P(\delta_{true} | \mathbf{X}) d\delta_{true}$$

For simplicity, a non-informative prior is assumed, and the biosimilarity index $BI(\mathbf{X})$ is approximated following Chow (2013; see also Shao & Chow, 2002):

$$BI(\mathbf{X}) \approx \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_L - t(1 - \alpha, df) \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} \right) - \Phi \left(\frac{(\mu_T - \mu_R) - \ln \theta_U + t(1 - \alpha, df) \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}}{\sqrt{2} \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_T^2}{n_T}}} \right)$$

The estimate of $BI(\mathbf{X})$, derived from a specific dataset \mathbf{X}_1 , is denoted as $BI(\mathbf{X}_1)$.

Similarly, estimated P_{RR} (\widehat{P}_{RR}) is obtained from computing $BI(\mathbf{X})$ using only the reference group data in \mathbf{X}_1 .

4.2.2 Estimation of P_{RR}

To estimate P_{RR} we leverage data collected from the reference group in the observed dataset. The process involves the following steps:

- **Step 1: Splitting the Reference Group Data**

The reference group data is randomly split into two non-overlapping subgroups, R_1 and R_2 , with sizes n_1 and n_2 respectively. The split ratio should maintain statistical power while minimizing the risk of bias due to small sample sizes.

- **Step 2: Computing Means and Variances**

The sample means ($\bar{X}_{R_1}, \bar{X}_{R_2}$) and variances ($S_{R_1}^2, S_{R_2}^2$) for the two subgroups are computed. These statistics represent the underlying variability and central tendency of the reference group data.

- **Step 3: Calculating the Biosimilarity Index within the Reference Group**

Using the calculated means and variances, the BI for R_1 versus R_2 is computed according to the predefined formula for BI. This value, denoted as P_{RR} , captures the degree of similarity between the two subgroups within the reference group.

It is important to note that a potential limitation of this approach arises when the sample size of the reference group is small, leading to challenges in adequately splitting the data into two subgroups while retaining statistical power. Strategies to mitigate this include:

- Increasing the overall reference group size during study design, if feasible.
- Applying resampling methods such as bootstrapping to improve the stability and reliability of estimation.
- Incorporating external data for the reference product to supplement the available information.

Once both $\widehat{BI}(\mathbf{X})$ and \widehat{P}_{RR} are obtained (estimation and statistical properties of $BI(\mathbf{X})$ can be found in Chapter 3), the estimate of the RBI (\widehat{RBI}) can be calculated, which serves as an index that accounts for variability inherent within the reference group.

4.3 Statistical Testing Methods for Relative Biosimilarity Index (RBI): A Comparative Framework

In this section, we present six statistical procedures for testing whether the Relative Biosimilarity Index (*RBI*) provides sufficient evidence to conclude biosimilarity. Specifically, we consider the hypothesis framework:

$$H_0: RBI \leq \delta \text{ (not biosimilar)} \quad \text{vs.} \quad H_1: RBI > \delta \text{ (biosimilar)},$$

where δ is a pre-specified decision threshold derived from the proposed function $\delta = \delta(CV)$, which will be discussed in Section 4.4.

The six methods differ in their inferential foundations and assumptions, such as how uncertainty in the numerator and denominator of *RBI* is addressed, whether distributional assumptions (e.g., normality or beta-distribution) are required, and whether resampling or regression modelling techniques are utilized. Each method provides different perspective for inference on *RBI* test, and the choice of method should be guided by how well it performs, in terms of type I error control and statistical power, under the proposed index (*RBI*) and decision threshold (δ). These operating characteristics will be evaluated through simulation in Section 4.5.

4.3.1 Fieller's Test

In this study, Fieller's method is considered to test whether the Relative Biosimilarity Index (*RBI*) exceeded a pre-specified threshold δ . While a standard Z-test could be applied by approximating the sampling distribution of *RBI* using the delta method, this approach assumes that the denominator, P_{RR} , is fixed or estimated with negligible uncertainty. However, in our context, both $BI(X)$ and P_{RR} are sample-based

estimators with non-negligible variability and potential dependence due to shared use of reference group data.

4.3.1.1 Test Statistic

This method is valid when the sample size is sufficiently large, or when $BI(X)$ and P_{RR} exhibit asymptotic joint normality. Fieller's method (Fieller, 1954) provides a theoretically valid test for such a ratio by evaluating the quadratic form:

$$T = \frac{(\widehat{BI(X)} - \delta \widehat{P_{RR}})^2}{\widehat{Var}(BI(X)) - 2\delta \cdot \widehat{Cov}(BI(X), P_{RR}) + \delta^2 \cdot \widehat{Var}(P_{RR})}$$

The variances and covariance can be either computed analytically (using the delta method) or estimated via nonparametric bootstrap resampling. Under the null hypothesis, the test statistic T approximately follows a chi-square distribution with 1 degree of freedom. The one-sided p-value can then be calculated as: $p = 1 - F_{\chi_1^2}(T)$, where $F_{\chi_1^2}(\cdot)$ is the cumulative distribution function of the chi-square distribution with 1 degree of freedom.

The null hypothesis will be rejected (i.e., conclude that the test product is biosimilar to the reference) if $p < \alpha$, where α is the pre-specified significance level. This approach appropriately accounts for uncertainty in both the numerator and denominator of the RBI . While Fieller's method, like the Z-test, relies on asymptotic normality, it more accurately accounts for uncertainty in both the numerator and denominator of the RBI . In small to moderate samples, the Z-test may underestimate the standard error of the ratio by ignoring denominator variability, leading to liberal inference. In contrast, Fieller's method models the ratio through a second-order

quadratic form, providing more conservative and reliable inference even when the normal approximation is only moderately accurate.

4.3.1.2 Variance and Covariance Estimation

4.3.1.2.1 Delta Method Estimation

According to Mai and Chow (2025), the variances of $BI(\mathbf{X})$ and P_{RR} can be approximated. Under the assumption that $BI(\mathbf{X})$ and P_{RR} are asymptotically normal, then their variances can be estimated through multivariate delta method. Since $BI(\mathbf{X})$ and P_{RR} are functions of μ_R and σ_R , then

$$Cov(BI(\mathbf{X}), P_{RR}) = \frac{\partial BI(\mathbf{X})}{\partial \mu_R} \cdot \frac{\partial P_{RR}}{\partial \mu_R} \cdot \frac{\sigma_R^2}{n_R} + \frac{\partial BI(\mathbf{X})}{\partial \sigma_R} \cdot \frac{\partial P_{RR}}{\partial \sigma_R} \cdot \frac{\sigma_R^2}{2n_R}$$

4.3.1.2.2 Bootstrap-Based Estimation

To relax the normality assumption underlying Fieller's method and provide a more robust inference framework, we employed a nonparametric bootstrap procedure to estimate the variance and covariance components required in the test statistic.

To compute this test statistic, test and reference data will be independently resampled (with replacement) K times. For each bootstrap sample k , $\widehat{BI(\mathbf{X})}^{(k)}$ and $\widehat{P_{RR}}^{(k)}$ will be computed and generated the set of bootstrap replicates $\left\{ \left(\widehat{BI(\mathbf{X})}^{(k)}, \widehat{P_{RR}}^{(k)} \right) \right\}_{k=1}^K$.

Variances and covariance can then be estimated:

$$\begin{aligned} \widehat{Var}(BI(\mathbf{X})) &= \frac{1}{K-1} \sum_{k=1}^K \left(\widehat{BI(\mathbf{X})}^{(k)} - \overline{BI(\mathbf{X})} \right)^2 \\ \widehat{Var}(P_{RR}) &= \frac{1}{K-1} \sum_{k=1}^K \left(\widehat{P_{RR}}^{(k)} - \overline{P_{RR}} \right)^2 \\ \widehat{Cov}(BI(\mathbf{X}), P_{RR}) &= \frac{1}{K-1} \sum_{k=1}^K \left(\widehat{BI(\mathbf{X})}^{(k)} - \overline{BI(\mathbf{X})} \right) \left(\widehat{P_{RR}}^{(k)} - \overline{P_{RR}} \right) \end{aligned}$$

This approach avoids direct reliance on the asymptotic joint normality of $(BI(\mathbf{X}), P_{RR})$, and instead uses the empirical distribution of these statistics to estimate their variability. This bootstrap-enhanced version of Fieller's method provides improved robustness in small to moderate sample sizes and accommodates potential non-normality or skewness in the sampling distribution of RBI .

4.3.2 Log Transformation

Since the biosimilarity index is bounded between 0 and 1 and is usually not too close to 0 or 1, a logit transformation may be considered. First is to apply the logit transformation to $BI(\mathbf{X})$ and P_{RR} . Then instead of working with $RBI = \frac{BI(\mathbf{X})}{P_{RR}}$, $RBI_{logit} = \text{logit}(BI(\mathbf{X})) - \text{logit}(P_{RR})$ will be computed, and the hypothesis will be changed accordingly:

$$H_0: RBI_{logit} \leq \text{logit}(\delta)$$

$$H_1: RBI_{logit} > \text{logit}(\delta)$$

A normality test (e.g. Shapiro-Wilk test) should be performed before conducting the test to ensure that the transformed data meets the normality assumption. If the normality holds, a standard Z-test can be performed with test statistic:

$$Z = \frac{RBI_{logit} - \text{logit}(\delta)}{\sqrt{\text{Var}(\text{logit}(BI(\mathbf{X}))) + \text{Var}(\text{logit}(P_{RR})) - 2\text{Cov}(\text{logit}(BI(\mathbf{X})), \text{logit}(P_{RR}))}}$$

The estimation of the variances of $BI(\mathbf{X})$ (and also P_{RR}) was discussed by Mai and Chow (2025), while the estimation of covariance follows the approach outlined earlier. Consequently, the estimated variances of logit transformed variables are given by:

$$\begin{aligned}\widehat{Var}(\text{logit}(BI(\mathbf{X}))) &\approx \left(\frac{1}{\widehat{BI}(\mathbf{X})(1 - \widehat{BI}(\mathbf{X}))}\right)^2 \text{Var}(\widehat{BI}(\mathbf{X})) \\ \widehat{Var}(\text{logit}(P_{RR})) &\approx \left(\frac{1}{\widehat{P}_{RR}(1 - \widehat{P}_{RR})}\right)^2 \text{Var}(\widehat{P}_{RR}) \\ \widehat{Cov}(\text{logit}(BI(\mathbf{X})), \text{logit}(P_{RR})) &\approx \left(\frac{1}{\widehat{BI}(\mathbf{X})(1 - \widehat{BI}(\mathbf{X}))}\right) \left(\frac{1}{\widehat{P}_{RR}(1 - \widehat{P}_{RR})}\right) \text{Cov}(\widehat{BI}(\mathbf{X}), \widehat{P}_{RR})\end{aligned}$$

The logit transformation offers computational efficiency and, when its underlying assumptions hold, leverages the power of parametric methods. However, its applicability is limited when $BI(\mathbf{X})$ is close to 0 or 1 or when the normality assumption remains violated despite the transformation.

4.3.3 Regression Inference with Bootstrap Sampling

To evaluate whether the observed Relative Biosimilarity Index (RBI) supports a claim of biosimilarity, we developed a set of inference procedures grounded in regression modelling. Given that RBI is a continuous quantity bounded within the open interval $(0, 1)$, beta regression is a natural framework for modelling both the magnitude and uncertainty of RBI estimates. This approach enables the testing of whether the mean RBI exceeds a threshold δ . The methodological foundation of beta regression is described by Ferrari and Cribari-Neto (2004), which provides the distributional assumptions and link functions used in modelling bounded responses. Moreover, the probability that RBI exceeds a threshold ($\Pr(RBI > \delta)$) can be assessed using a Bernoulli or logistic regression model applied to bootstrap replicates of the RBI .

4.3.3.1 Testing Mean RBI

Again, the test and reference data will be independently resampled (with replacement) K times. Let $RBI^{(k)} \in (0,1)$ denote the RBI calculated from the k^{th} bootstrap sample, with $k = 1, \dots, K$. Each $RBI^{(k)}$ value is modelled as a draw from the Beta distribution:

$$RBI^{(k)} \sim Beta(\mu, \phi),$$

where $\mu \in (0,1)$ is the mean and $\phi > 0$ is the precision parameter. A logit link function can be used to model the mean:

$$logit(\mu) = \beta_0$$

The hypotheses are formed as:

$$H_0: \mu \leq \delta \quad \text{and} \quad H_1: \mu > \delta,$$

which is equivalent to:

$$H_0: \beta_0 \leq \log\left(\frac{\delta}{1-\delta}\right).$$

Then construct a Wald statistic:

$$Z = \frac{\hat{\beta}_0 - \log(\delta/(1-\delta))}{SE(\hat{\beta}_0)},$$

where $\hat{\beta}_0$ is the estimated intercept and $SE(\hat{\beta}_0)$ is the standard error. The one-sided p-value is then $p = 1 - \Phi(Z)$, and the null hypothesis can be rejected when $p < \alpha$, which indicates that the mean RBI is significantly greater than δ , supporting the claim of biosimilarity.

4.3.3.2 Testing $Pr(RBI > \delta) \geq 1 - \alpha$

To reflect regulatory interpretation of biosimilarity as a high-probability event, we can also test on whether the probability that RBI exceeds δ is at least $1 - \alpha$. For each bootstrap sample, we define:

$$z^{(k)} = \mathbb{1}(RBI^{(k)} > \delta),$$

and modelled:

$$z^{(b)} \sim \text{Bernoulli}(p),$$

where $p = Pr(RBI > \delta)$ is the probability of *RBI* greater than the threshold δ .

The hypotheses are formulated as:

$$H_0: p \leq 1 - \alpha \quad \text{and} \quad H_1: p > 1 - \alpha.$$

Then the proportion:

$$\hat{p} = \frac{1}{K} \sum_{k=1}^K z^{(k)}$$

can be tested using the following two approaches:

(a) One-sample binomial proportion test

Suppose the total bootstrap samples K is sufficiently large, the test statistic would be:

$$Z = \frac{\hat{p} - (1 - \alpha)}{\sqrt{(1 - \alpha)\alpha/K}}$$

and the corresponding one-sided p-value can be computed as $p = 1 - \Phi(Z)$.

Alternatively, the p-value can be computed using an **exact one-sided binomial test**. Under the null hypothesis that the probability of *RBI* exceeding δ is at most $1 - \alpha$,

the number of observed exceedances among K replicates follow a Binomial distribution with parameters K and $1 - \alpha$. The p-value is then given by:

$$p = \Pr(\text{Binomial}(K, 1 - \alpha) \geq K \cdot \hat{p}),$$

which represents the probability of observing as many or more RBI values exceeding δ as in the data, if the true probability of exceedance were only $1 - \alpha$. A small p-value indicates strong evidence that $\Pr(RBI > \delta) > 1 - \alpha$, thus supporting the claim of biosimilarity.

(b) Logistic regression (if covariates are included)

In this method, the model would be:

$$\text{logit}(p) = \eta_0,$$

and the hypotheses are:

$$H_0: \eta_0 \leq \log\left(\frac{1 - \alpha}{\alpha}\right) \quad \text{and} \quad H_1: \eta_0 > \log\left(\frac{1 - \alpha}{\alpha}\right).$$

A Wald statistic can then be calculated as:

$$Z = \frac{\hat{\eta}_0 - \log((1 - \alpha)/\alpha)}{SE(\hat{\eta}_0)}.$$

The null hypothesis can be rejected when the one-sided p-value is less than pre-specified significance level α , which means there is strong evidence that $\Pr(RBI > \delta) > 1 - \alpha$, therefore supports the claim of biosimilarity.

The p-value from the logistic regression Wald test represents the strength of evidence against the null hypothesis that the probability of RBI exceeding the threshold δ is at most $1 - \alpha$. A small p-value indicates strong support that $\Pr(RBI > \delta) > 1 - \alpha$, and thus, that the biosimilarity criterion is met with the desired level of confidence.

4.3.4 Comparison of Testing Methods

Several methods have been proposed in this study to test the hypothesis $H_0: RBI \leq \delta$, including Fieller's method, logit-transformed Z-tests, and regression modelling. While these methods differ in their underlying statistical assumptions and inference targets, they can all be used with a common decision threshold δ , provided that appropriate scale alignment is maintained.

Table 22: Comparative summary of six statistical methods.

Method	Inferential Basis	Key Assumptions	Strengths	Limitations
Fieller's Method	Ratio-based confidence interval and quadratic inequality	Approximate joint normality of BI and P_{RR} ; large sample size	Explicitly accounts for variability in both numerator and denominator; widely used in bioequivalence	Unstable when denominator (P_{RR}) is small or highly variable; requires large-sample normality
Logit Transformation	Z-test on logit-transformed	Approximate normality of $logit(BI)$ and $logit(P_{RR})$;	Straightforward implementation; variance estimable on	Requires RBI components in $(0,1)$; sensitive to

	RBI components	moderate sample size	transformed scale	skewness; may break down if values are near boundaries
Beta Regression	Parametric regression modelling of <i>RBI</i> as a Beta-distributed response	$RBI \in (0,1)$; Beta-distribution appropriate; independent bootstrap replicates	Suitable for bounded, continuous outcomes; supports direct hypothesis on mean <i>RBI</i>	Requires bootstrap samples; may be sensitive to boundary values; limited interpretability of coefficients
Binomial Approximation	Z-test on proportion of bootstrap <i>RBI</i> s exceeding δ	Large K (replicates); CLT approximation for Binomial	Easy to implement; interprets $RBI > \delta$ as a probabilistic event	May lack accuracy with few exceedances; approximation breaks down when p is near 0 or 1

<p>Binomial Exact Test</p>	<p>Exact one-sided test on number of $RBI > \delta$ exceedances</p>	<p>None (non-parametric); requires binomial modelling of bootstrap exceedances</p>	<p>Distribution-free; valid under small sample conditions</p>	<p>Discrete p-values; limited resolution; may lack power</p>
<p>Logistic Regression</p>	<p>Wald test on intercept from logistic regression of $RBI > \delta$</p>	<p>Bernoulli model; logit link; no quasi-complete separation; sufficient exceedances</p>	<p>Enables covariate adjustment; interprets biosimilarity as high probability event</p>	<p>May suffer from separation when few $RBI > \delta$; unstable estimates with rare events; large K required</p>

Each method provides a different perspective on the testing problem. Fieller's method is ratio-based and useful for direct comparison of RBI to δ . Logit-based tests leverage transformation for variance stabilization but require scale transformation of δ . Regression allows modelling of the expected RBI or its exceedance probability, with greater flexibility for incorporating covariates or modeling heteroscedasticity.

The decision threshold δ is selected a priori based on regulatory equivalence bounds and prior assumptions about between-subject variability (e.g., CV). This selection is agnostic to the specific testing procedure and is intended to be broadly applicable. However, care must be taken to ensure that the test statistic and the interpretation of δ are aligned on the same scale. For instance, when logit-transformation is applied to RBI , the corresponding threshold must be transformed to $logit(\delta)$; in probability-based beta regression, δ is treated as a reference for exceedance probability rather than a direct comparator. While the δ value itself is fixed, the way it is used in each test may differ slightly depending on the structure and scale of the model.

4.4 Method for Selecting δ in Testing the Relative Biosimilarity Index (RBI)

The selection of the threshold δ in testing the Relative Biosimilarity Index (RBI) is critical to ensuring both compliance with regulatory standards and clinical relevance. Regulatory agencies such as the FDA define a biosimilar as a product that is “highly similar to the reference product notwithstanding minor differences in clinically inactive components,” and that has “no clinically meaningful differences in terms of safety, purity, and potency” (FDA, 2015). The threshold δ must therefore reflect these regulatory expectations while incorporating statistical variability inherent in the data. Below, we propose a structured method for determining δ that integrates regulatory equivalence criteria and variability adjustment.

4.4.1 Method

To ensure δ is consistent with regulatory standards and practical feasibility, the following steps are proposed:

- **Step 1: Define Key Parameters**

The threshold δ represents the minimum acceptable *RBI* value above which biosimilarity is concluded. Regulatory agencies commonly define equivalence margins for pharmacokinetic (PK) parameters (e.g., (0.8,1.25) for the geometric mean ratio). These bounds serve as the basis for determining δ , ensuring it aligns with clinical expectations. The baseline threshold δ_0 is derived from these equivalence margins and represents the default *RBI* threshold without adjustments for variability.

- **Step 2: Translate Regulatory Benchmarks into δ**

To convert these equivalence bounds into a baseline *RBI* threshold, we use the relation:

$$\delta = \frac{1}{1 + M'}$$

where M is the maximum allowed deviation (e.g., e.g., $M = 0.25$ for the bounds (0.8, 1.25)). This formula ensures that the *RBI* reflects a lower bound for acceptable similarity, consistent with regulatory guidelines (see EMA, 2014 and FDA, 2022).

- **Step 3: Incorporate Variability into δ**

To account for variability in the reference product, we introduce an adjustment based on the coefficient of variation (CV). The adjusted threshold is defined as:

$$\delta = \delta(CV) = \frac{\delta_0}{1 + \gamma \cdot CV},$$

where:

- δ_0 is the baseline threshold from Step 2,
- CV is the observed coefficient of variation of the reference product, where s^2 is the sample variance of the reference data (log-transformed if using PK).

$$CV = \sqrt{e^{s^2} - 1}$$

- γ is a tuning parameter reflecting tolerance for variability.

This adjustment ensures the threshold accounts for the inflation of variability under biologics and aligns with regulatory standards. The idea of adjusting statistical thresholds for variability is conceptually similar to scaled average bioequivalence (SABE) approaches used for highly variable drugs, where equivalence margins are adapted based on observed variability in the reference product (FDA, 2001; Chow, 2014; Tothfalusi & Endrenyi, 2016; Haidar et al., 2008; Davit et al., 2012). In particular, methods like scaled average bioequivalence (SABE) adjust the margins in proportion to the reference product's within-subject variation, ensuring that drugs with higher inherent variability are not unfairly penalized. In the setting of this

paper, between-subject variability is the most relevant source of uncertainty, as a parallel design is used. Therefore, we adjust the *RBI* threshold δ using the coefficient of variation of the reference group, reflecting population-level variability in a manner analogous to SABE's use of intra-subject variability.

- **Step 4: Perform Sensitivity Analysis**

To ensure the robustness of the selected threshold $\delta(CV)$, we recommend a comprehensive sensitivity analysis:

- Vary *CV* across plausible ranges (e.g., 0.1, 0.15, 0.2),
- Vary sample sizes across plausible ranges,
- Explore different values of γ (e.g., 0.2, 0.3, 0.4),
- Evaluate the resulting impact on type I error and power,

This step helps the selection of threshold δ so that it balances regulatory stringency with practical considerations, while ensuring adequate statistical power ($\geq 80\%$) and controlled Type I error ($\leq 5\%$).

4.4.2 Choosing the Tuning Parameter γ

The parameter γ in the adjustment formula plays a crucial role in determining how much the baseline threshold δ_0 is adjusted based on the variability (*CV*) of the reference product.

The selection of γ should balance regulatory stringency, statistical rigor, and feasibility:

- Regulatory Alignment: Examine precedents in biosimilar product approvals and relevant guidance to identify implicit tolerance levels for variability.
- Statistical Considerations: Choose γ to maintain type I error control ($\leq 5\%$) and adequate power ($\geq 80\%$). Simulation studies (e.g., via Monte Carlo or bootstrap) can be used to assess the operating characteristics of different γ values.
- Example Ranges:
 - $\gamma = 0.1$ to 0.3 : suitable when variability is low and regulatory stringency is high.
 - $\gamma = 0.3$ to 0.5 : appropriate when more flexibility is acceptable, especially for biologics with higher variability.

This framework should provide a reproducible and flexible method for selecting the biosimilarity threshold δ used in testing the Relative Biosimilarity Index. By integrating regulatory equivalence criteria, adjustments for variability, and sensitivity analyses, the approach supports statistically valid and clinically interpretable biosimilarity claims. It is adaptable to diverse trial settings and promotes consistency in decision-making across studies.

4.5 Simulation Study

The simulation study was designed to support three key objectives in the development and application of the Relative Biosimilarity Index (*RBI*) framework. First, we conducted simulations to provide guidance on selecting the tuning parameter γ in

the threshold function $\delta = \delta(CV) = \delta_0/(1 + \gamma \cdot CV)$, which adaptively defines the similarity threshold based on the coefficient of variation (CV). Second, we evaluated how sample size and variability jointly influence the operating characteristics of the RBI-based test, specifically empirical type I error and power, offering practical guidance for sample size determination under specified design settings (i.e., given CV and δ). Third, we compared the performance of six statistical testing procedures applied to the RBI, using the decision threshold δ derived from the proposed function, to assess their suitability in terms of error control and inferential robustness.

Datasets were simulated under both the null hypothesis (non-biosimilarity) and the alternative hypothesis (biosimilarity), across a range of CV values and sample sizes. The RBI was computed for each dataset and compared to the pre-specified threshold δ , and empirical type I error and power were estimated based on the proportion of rejections under H_0 and H_1 , respectively. Where applicable, the simulations were tailored to align with the specific assumptions of each testing method. This simulation framework should provide practical guidance for calibrating δ , selecting appropriate sample sizes, and choosing robust testing procedures in the context of biosimilarity assessment in our proposed framework.

Before conducting simulation, the following properties are first considered to ensure that δ function can be deemed valid in the context of biosimilarity testing based on the Relative Biosimilarity Index (RBI):

(i) Type I Error Control

A valid δ function must ensure that the test controls the familywise Type I error rate at or below a nominal significance level (typically $\alpha = 0.05$) under the null hypothesis. Specifically, when the test and reference products are not biosimilar, the probability of incorrectly declaring biosimilarity should satisfy:

$$\Pr(\widehat{RBI} > \delta | H_0) \leq \alpha.$$

This property is evaluated empirically in the simulation by generating data under H_0 (e.g., with a log mean difference exceeding the regulatory equivalence margin) and computing the proportion of simulations in which RBI exceeds the threshold δ .

(ii) Power Adequacy

The function δ should not be overly conservative. Under the alternative hypothesis, when the test and reference products are truly biosimilar, the probability of rejecting H_0 (i.e., correctly concluding biosimilarity) should be high, for instance $power \geq 0.8$. Thus, the function must achieve:

$$\Pr(\widehat{RBI} > \delta | H_1) \geq 0.8.$$

(iii) Interpretability and Clinical Alignment

The function should produce δ values that lie in the interval $(0, 1)$ and remain interpretable in the context of biosimilarity. Ideally, the function is derived from regulatory equivalence margins (e.g., $0.8 - 1.25$ for PK endpoints) and scaled to reflect study-specific variability, while retaining a meaningful link to clinically acceptable similarity levels.

(iv) Monotonicity and Continuity

To ensure logical and statistical consistency, the function δ should be continuous and monotonic with respect to the coefficient of variation. For instance, a decreasing function like $\delta(CV) = \delta_0 / (1 + \gamma \cdot CV)$ intuitively reflects that achieving a high *RBI* becomes more difficult under high variability, and therefore appropriately relaxes the threshold.

4.5.1 Determination of γ

To evaluate the performance of the *RBI*-based testing procedure under varying design conditions, we conducted a series of simulations across a range of coefficient of variation (*CV*) values and tuning parameters γ . The threshold δ was determined by the function $\delta(CV) = \delta_0 / (1 + \gamma \cdot CV)$, with δ_0 fixed at 0.8. The resulting empirical power and type I error were estimated under both the null hypothesis (non-biosimilarity) and the alternative hypothesis (biosimilarity). Figures 1–4 illustrate the relationships among power, type I error, *CV*, and γ .

The following figures are based on 10,000 simulation replicates conducted under the specified parameter settings: (1) the mean difference under the null hypothesis is set to 0.4; (2) the mean difference under the alternative hypothesis is set to 0.22; (3) sample size per treatment is 50; and (4) the standard deviation is assumed to be equal to the coefficient of variation (*CV*), and the similarity threshold $\delta = \delta(CV)$.

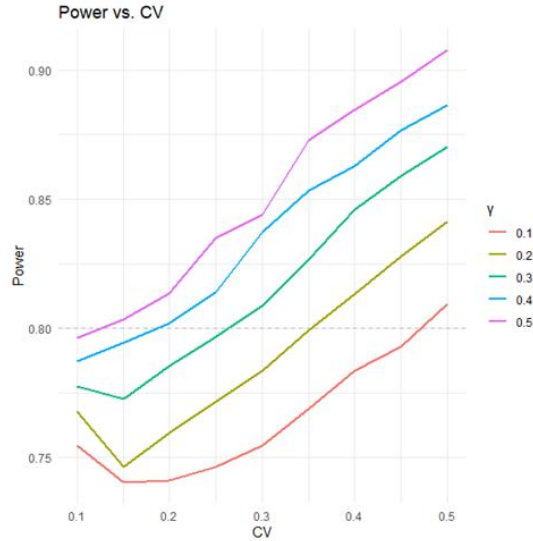


Figure 13: Power vs. CV for different γ values.

This plot shows the power of the RBI-based biosimilarity test as a function of the coefficient of variation (CV), stratified by the tuning parameter γ used in the threshold function $\delta(CV) = \delta_0/(1 + \gamma \cdot CV)$. As the CV increases, higher γ will generate higher power, indicating that a more aggressive decrease in δ helps mitigate the loss in power caused by increasing variability.

Figure 13 shows how statistical power varies with the coefficient of variation (CV), stratified by different values of the tuning parameter γ . For smaller γ values (e.g., $\gamma = 0.1$), power initially declines as CV increases from 0.1 to around 0.2, then gradually recovers as CV increases further, forming a U-shaped relationship. In contrast, for larger values of γ , power increases more steadily across the full CV range. This pattern reflects the interaction between the widening of the RBI distribution under higher variability and the adaptively decreasing threshold $\delta(CV)$. At low CV levels, the RBI distribution is concentrated above δ , leading to high power. As CV increases, the distribution becomes wider and fewer RBI values exceed the threshold, reducing power. However, at higher

CV s, the threshold δ becomes increasingly lenient due to its dependence on γ and CV , compensating for the wider distribution and leading to a partial or full recovery in power. This recovery is more prominent for larger γ values, due to a more aggressive reduction of δ as CV increases.

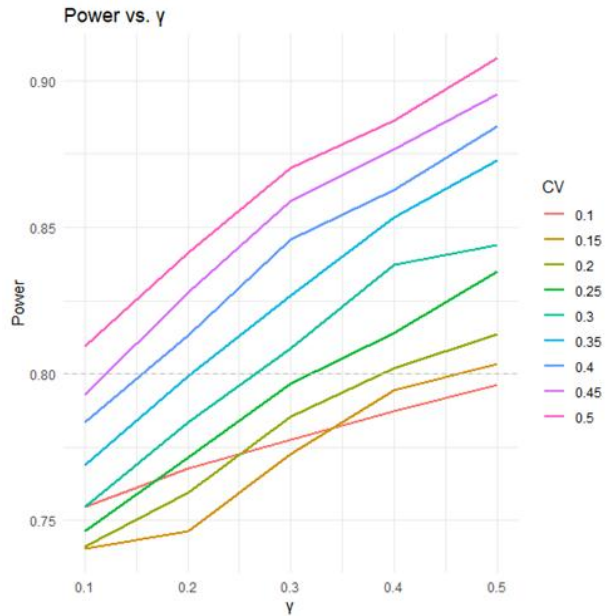


Figure 14: Power vs. γ for different CV levels.

Figure 14 explores the effect of γ on power, stratified by CV . Across all levels of CV , increasing γ leads to an increase in power. This is because higher values of γ reduce δ more substantially, making it easier for the RBI to exceed the threshold. Notably, power is consistently higher for greater CV values across all values of γ , suggesting that the combination of a more lenient threshold and greater spread in the RBI distribution under high variability increases the chance of detecting biosimilarity. This reinforces that larger values of γ are more effective in compensating for the inflation of variability.

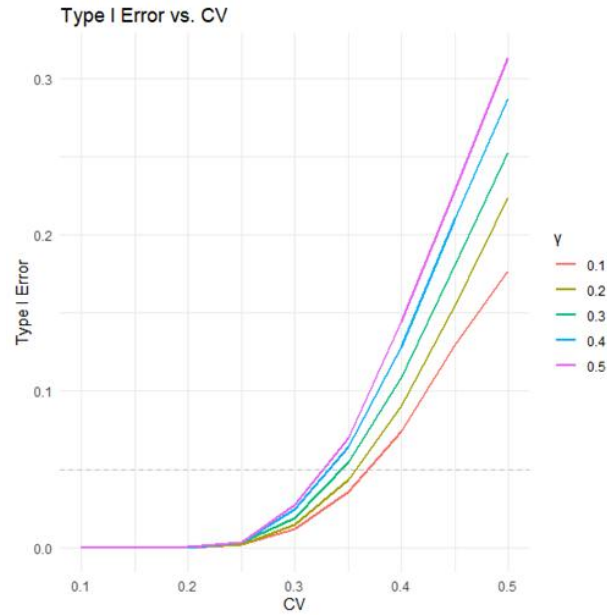


Figure 15: Type I error vs. CV for different γ values.

Figure 15 presents the type I error across CV levels for each value of γ . At low CV (e.g., $CV < 0.25$), type I error remains well-controlled for all γ values. However, as CV increases beyond this point, type I error begins to rise and eventually exceeds the nominal 0.05 threshold, particularly when $\gamma \geq 0.32$. This indicates that while increasing γ improves power, it can compromise the validity of the test by allowing null distributions of RBI to frequently exceed overly lenient δ thresholds. The inflation of type I error is more pronounced under higher CV s, where the relaxation of δ is most substantial.

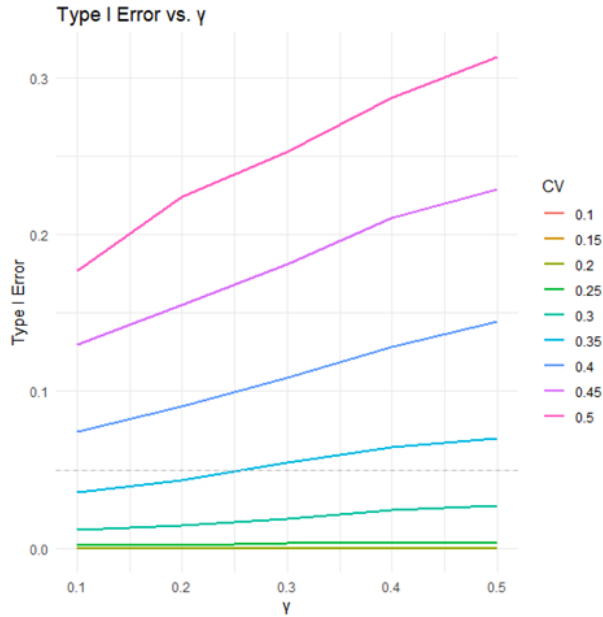


Figure 16: Type I error vs. γ for different CV levels.

Figure 16 provides a complementary view by illustrating how type I error increases with γ , stratified by CV . For very low CV values (e.g., $CV = 0.1$), type I error remains negligible regardless of the γ value. However, for high CV values (e.g., $CV = 0.4 - 0.5$), type I error grows sharply as γ increases and generally exceeds the nominal level across all γ values. For moderate CV values (e.g., $CV = 0.35$), the type I error is typically greater than 0.05 once γ exceeds approximately 0.3. This reinforces the finding that while a higher γ enhances power, it also introduces the risk of elevated false positive rates, especially in the presence of high variability.

Together, these results offer guidance on how to select the tuning parameter γ when applying the RBI-based biosimilarity test with an adaptively determined threshold δ . Higher γ values enhance statistical power, especially under conditions of high variability, but at the cost of inflated type I error. Conversely, smaller values of γ preserve type I error control but may lead to inadequate power. These findings suggest

that moderate values of γ , such as those in the range of 0.2 to 0.3, may offer a reasonable trade-off between sensitivity and specificity, yielding reliable performance across practical ranges of CV (e.g., 0.1 – 0.35) commonly observed in biosimilarity studies. These insights support the selection of γ and planning of trial designs that appropriately balance regulatory rigor with statistical efficiency.

4.5.2 Simulation-Based Evaluation of Sample Size and Type I Error Control

To guide the practical implementation of the RBI-based biosimilarity testing procedure, we conducted a comprehensive simulation study to evaluate its operating characteristics under various conditions of between-subject variability and sample size. Specifically, we examined how the choice of threshold $\delta = \delta(CV)$, in conjunction with sample size and coefficient of variation (CV), affects the type I error control and statistical power of the RBI-based test.

4.5.2.1 Power-Based Sample Size Guidance

The power heatmap (e.g. Figure 17) evaluates the proportion of rejections under the alternative hypothesis (true biosimilarity), illustrating how both sample size and δ influence the ability to detect biosimilarity when it exists. In addition to the graphical representations, design tables that summarize the minimum sample size required to achieve at least 80% power for each threshold $\delta(CV)$, at a fixed CV , can be constructed (e.g. Table 23). These tools provide practical guidance for selecting sample sizes that ensure sufficient statistical power of the RBI-based biosimilarity test.

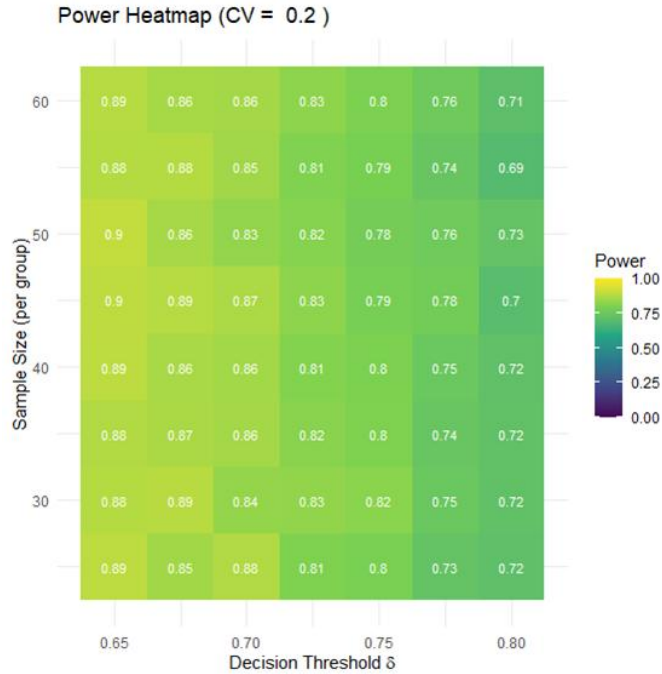


Figure 17: Power with different sample sizes and δ at $CV = 0.2$.

Table 23: Empirical minimum sample size for achieving sufficient power (80%) for varying δ , with fixed coefficient of variation ($CV = 0.20$) and mean difference ($mean_{diff} = 0.1$).

δ	Minimum Sample Size (per group)	Power
0.65	25	0.893
0.675	25	0.854
0.7	25	0.876
0.725	25	0.806
0.75	30	0.82

4.5.2.2 Type I Error Evaluation and Sample Size Validation

Separately, type I error control was evaluated under the null hypothesis (non-biosimilarity) for each sample size, assuming the expected coefficient of variation (CV)

for the reference group and the corresponding decision threshold δ . Figure 18 shows how the empirical type I error decreases as the sample size increases, supporting the selection of appropriate sample sizes to maintain type I error control. The resulting design table (e.g., Table 24) summarizes the empirical type I error rates across a range of sample sizes. This evaluation supports the identification of sample size settings that maintain type I error below the nominal level (e.g., 5%) for the specified threshold. These simulation results provide additional validation for the sample size recommendations derived from the power considerations discussed in the previous section.

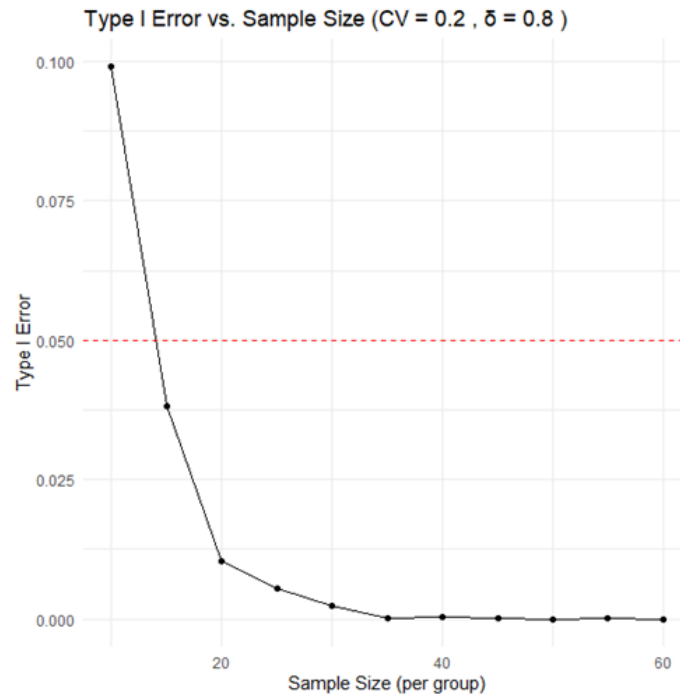


Figure 18: Type I error rate as a function of sample size per group for fixed coefficient of variation ($CV = 0.20$) and decision threshold ($\delta = 0.8$).

Table 24: Empirical type I error rates under the null hypothesis for varying sample sizes, with fixed coefficient of variation ($CV = 0.20$) and mean difference (0.4).

δ	Sample Size (per group)	Type I Error
0.8	10	0.0990
0.8	15	0.0382
0.8	20	0.0104
0.8	25	0.0054
0.8	30	0.0002
0.8	35	0.0004
0.8	40	0.0002
0.8	45	0.0000

Together, the power and type I error analyses provide comprehensive insights into the selection of appropriate sample sizes after the decision threshold δ is determined, thereby supporting the planning and efficient design of biosimilarity studies using the proposed RBI-based framework.

4.5.3 Method-Specific Performance Evaluation under Selected Threshold

Sections 4.5.1 and 4.5.2 focused on selecting the threshold function $\delta(CV)$ and determining minimum sample sizes needed to achieve desired type I error control and statistical power, based on direct comparison of the empirical *RBI* to δ . However, subsequent simulations revealed that the required sample size to attain acceptable performance may vary substantially depending on the statistical testing method used. To investigate this observation, we conducted a method-specific performance evaluation under a fixed decision threshold $\delta(CV)$, comparing six inference procedures: Fieller's method, logit transformation, beta regression, binomial-Z test, binomial exact test, and

logistic regression (Wald test). Each method was evaluated using simulations designed to meet its respective model assumptions, with empirical type I error and power assessed across a range of sample sizes. This section demonstrates that, although all methods test the same hypothesis using the RBI and the same threshold, their operating characteristics, especially under small-to-moderate sample sizes, differ significantly. These differences underscore the importance of selecting an inference method that is well-aligned with both the study design and desired performance metrics.

Each method was applied to simulated datasets under both the null hypothesis (non-biosimilarity) and the alternative hypothesis (biosimilarity), using a fixed threshold $\delta(CV)$ determined by the proposed formula. Sample sizes were systematically varied to evaluate each method's empirical type I error and statistical power under realistic conditions.

Figures 19 and 20 present the empirical type I error rates and statistical power curves, respectively, for each of the six testing methods across a range of sample sizes. To ensure a fair and valid comparison, the data for each method were simulated to match the assumptions underlying its inferential framework, for example joint normality for Fieller's, beta-distributed RBI values for beta regression.

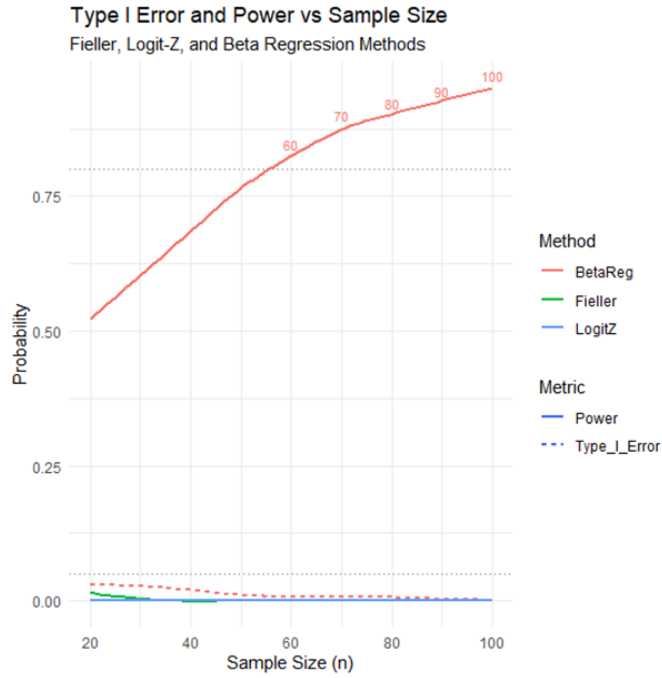


Figure 19: Type I error and power across varying sample size for Fieller’s method, beta regression and logit transformation tests.

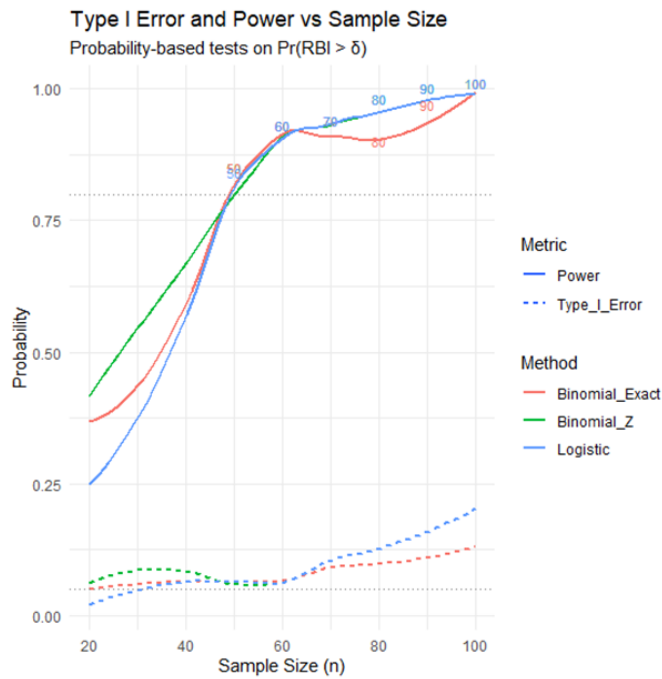


Figure 20: Type I error and power across varying sample size for probability-based tests.

Among all methods, beta regression demonstrated the most robust and balanced performance, consistently maintaining appropriate type I error control while achieving desirable statistical power (≥ 0.8) for sample sizes (per arm) larger than approximately 55. In contrast, Fieller's method and the logit transformation approach exhibited unacceptably low power, even as sample size increased, suggesting limitations in their applicability under the given conditions. The three probability-based methods: binomial approximation, binomial exact test, and logistic regression, achieved acceptable power for sample size (per arm) larger than 50, but showed slightly inflated and fluctuated type I error rates.

A possible explanation is that RBI is a ratio of two noisy estimators. Ratio-based and transformation-based tests rely on large-sample normal approximations that can be fragile when P_{RR} is highly variable or when values lie near the boundaries of the parameter space. In contrast, the regression procedures operate on the natural bounded (0,1) scale of RBI, model the associated mean–variance relationship, and use bootstrap replicates to propagate the joint uncertainty in \widehat{BI} and \widehat{P}_{RR} . This combination reduces variability of the test statistic and delivers type I error control closer to the nominal level, which explains the comparatively robust performance observed in our simulations.

Among the regression options, beta regression is the most stable because it analyses RBI as a continuous outcome on (0,1) with both a mean and a precision parameter. This preserves information by using the full RBI values rather than dichotomizing them. Exceedance-based approaches replace each replicate with the indicator $\mathbb{1}(\text{RBI}^{(k)} > \delta)$, which discards the distance from δ and introduces discreteness

and potential rare-event issues. Consequently, beta regression achieved more stable type I error control across the scenarios considered.

These results highlight the practical implications of method selection in RBI-based biosimilarity testing. Although the threshold $\delta = \delta(CV)$ is defined independently of the testing procedure, selecting a robust and well-performing method is essential to ensure valid inference when applying the proposed RBI framework. Our results suggest that method choice should be guided by its ability to maintain type I error control and adequate power under the chosen similarity threshold.

4.6 Discussion and Concluding Remarks

In the study discussed in this chapter, we proposed and systematically evaluated a newly defined metric—the Relative Biosimilarity Index (RBI)—as a flexible and interpretable framework for biosimilarity assessment. The RBI is constructed as the ratio between the estimated biosimilarity index comparing the test to the reference product (BI) and the estimated self-similarity index of the reference product (P_{RR}). By incorporating P_{RR} into the denominator, the RBI directly accounts for the variability inherent in the reference product, which is a critical factor in biosimilarity evaluation but often overlooked in conventional fixed-margin approaches.

To test whether the RBI exceeds a pre-specified threshold δ , we developed and examined six statistical inference procedures under the hypothesis framework $H_0: RBI \leq \delta$ versus $H_1: RBI > \delta$. These include Fieller’s method, which accounts for the joint variability of the numerator and denominator; a logit transformation approach, which linearizes the ratio for inference under approximate normality; a beta regression model

that leverages the bounded nature of *RBI*; and three bootstrap-based probability tests, including the binomial approximation, binomial exact test, and logistic regression (Wald test). Each method differs in assumptions, inferential strategies, and robustness to data features.

A systematic threshold selection method was proposed using $\delta(CV) = \delta_0 / (1 + \gamma \cdot CV)$, where δ_0 is based on regulatory equivalence bounds, and γ is a tuning parameter reflecting the degree of adjustment for between-subject variability. This approach provides a principled, interpretable method for adjusting the similarity threshold to preserve type I error control and maintain statistical power under different levels of variability.

To support implementation, we generated empirical heatmaps and design tables for selecting sample size under various levels of *CV* and δ thresholds. These visual and tabular tools offer a data-driven basis for planning studies with adequate power and proper type I error control, facilitating the application of the *RBI*-based framework in practice.

Simulation studies were also conducted to evaluate the performance of these testing procedures under the proposed threshold function. The results revealed that while all methods are theoretically valid under certain conditions, their empirical performance varies. Beta regression demonstrated strong control of type I error with robust power, with moderate sample sizes (approximately over 55 per arm). In contrast, Fieller's method and logit-based testing displayed limited power, despite assumptions being satisfied in the simulations. The three probability-based methods exhibited

moderate to good power, though with mild inflation in type I error rates. These findings suggest that although the threshold δ is defined independently of the testing procedure, method selection is critical. Robust inference on *RBI* requires selecting a procedure that performs well with respect to both type I error control and power, given the practical constraints of biosimilar trial design.

It is important to note that even when the data are simulated to satisfy all model assumptions (e.g., normality for Fieller's method, logit link for logistic regression, beta-distributed *RBI* for beta regression), some methods still require larger sample sizes to achieve acceptable performance. This phenomenon stems from the reliance on asymptotic approximations in methods such as Fieller's test and Wald-based regression models, which may not perform adequately in small samples. Additionally, the use of ratio-based estimators (as in *RBI*) inherently introduces higher variance, especially in small samples, further amplifying instability in the test statistics. For boundary-sensitive metrics like *RBI*, inference becomes fragile when values cluster near 0 or 1. Even bootstrap-based methods such as beta regression or binomial tests require stable resampling distributions that are more reliable with moderate-to-large sample sizes. These findings suggest that satisfying model assumptions alone is not sufficient for reliable inference at small sample size, further underscoring the need to complement threshold selection with careful choice of testing method and appropriate sample size planning.

While the current approach uses fixed thresholds based on prior assumptions of variability, future work could incorporate Bayesian extensions. Specifically, variability

parameters such as CV could be treated as random quantities with prior distributions, and $\delta(CV)$ could be integrated over their posterior distributions to yield an unconditional threshold. This would allow greater flexibility in accounting for uncertainty and variability in real-world trial settings. Additionally, extending the *RBI* framework to accommodate multivariate endpoints, repeated measures, or longitudinal outcomes may enhance its applicability to complex biosimilarity evaluations. Validation using real clinical data and case studies will also be crucial to establish the utility and regulatory relevance of the proposed methods.

In conclusion, the Relative Biosimilarity Index, along with the proposed threshold selection strategy and accompanying simulation-based evaluation, provides a rigorous and adaptable framework for modern biosimilarity testing. While several inference procedures were explored, the goal is not to offer interchangeable alternatives but rather to identify the appropriate method that performs reliably in controlling type I error and maintaining adequate power when used in conjunction with the *RBI* and its associated threshold δ . The simulation results underscore the importance of method selection tailored to the characteristics of the biosimilarity study, reinforcing the practical utility of the proposed framework in guiding both regulatory decisions and study design.

5. Conclusion and Future Directions

This thesis has focused on advancing the statistical methodology of two stage seamless adaptive designs and their application to biosimilar drug development.

Motivated by the growing complexity of modern clinical trials and the need for greater

efficiency in regulatory science, the research addressed important methodological gaps that arise when designs must adapt to changes in populations, endpoints and decision-making thresholds. Across three major contributions, analysing two stage seamless adaptive designs with population and endpoint shifts (Chapter 2), proposing an innovative PK–clinical two stage design for biosimilar development (Chapter 3), and developing the Relative Biosimilarity Index with a structured framework for similarity threshold selection (Chapter 4), the work provides a coherent set of advances that link statistical theory with pressing practical problems in drug development.

5.1 Integration of Contributions

Although each chapter presents its own methodological focus, the three core contributions are interconnected and should be understood as parts of a broader narrative. In Chapter 2, a general methodological framework was developed for analysing two stage seamless adaptive designs in which study endpoints and patient populations may differ between stages while the overarching study objective remains the same. This methodological development is motivated by practical realities. Protocol amendments and disease progression can lead to shifts in the study population, and in many trials, it is common to employ a short-term surrogate marker at an early stage and a long-term primary endpoint at a later stage. The proposed framework provided a systematic way to link information across stages under these conditions, ensuring type I error control and statistical power, while also delivering interpretable conclusions on treatment efficacy in the shifted population.

Chapter 3 built directly on this foundation by applying adaptive design principles to the context of biosimilar drug development. Biosimilar programs, unlike those for novel drugs, must establish similarity rather than superiority, and they must do so across multiple evidentiary domains, including analytical characterization, pharmacokinetics, and clinical outcomes. Traditionally, these domains are evaluated in a sequential stepwise fashion, which can be time-consuming and resource intensive. By proposing a two-stage seamless PK–clinical adaptive design, Chapter 3 illustrated how the general principles of adaptive methodology can be tailored to meet the unique needs of biosimilar development. Under the Fundamental Biosimilarity Assumption that PK similarity predicts clinical similarity, the design enabled early decision-making at the conclusion of stage one. This created the possibility of determining biosimilarity more rapidly when PK evidence was sufficiently strong, thereby reducing the burden of duplicative studies and shortening the overall development timeline.

Chapter 4 then advanced this line of work by addressing a limitation of the biosimilarity index framework that underpinned the design in Chapter 3. While the BI provides a robust and flexible measure for assessing similarity, its application in practice requires a clear specification of decision thresholds. Without principled guidance for threshold selection, conclusions may be either overly lenient or overly stringent, with implications for both statistical validity and practical feasibility. To address this, Chapter 4 introduced the Relative Biosimilarity Index (RBI), which normalises the BI against reference-to-reference variability, thereby providing a more stable and interpretable foundation for inference. In addition, a structured methodology for selecting the

decision threshold δ was proposed, drawing on equivalence margins and adjustments for variability. Simulation studies showed how threshold choice, sample size, and inference procedure affect type I error and power, thereby offering practical guidance for implementation.

Together, these contributions represent a coherent progression. Chapter 2 established general principles for valid inference under complex trial adaptations, Chapter 3 translated those principles into a concrete application for biosimilar development, and Chapter 4 refined the inferential framework by formalising threshold determination. The trajectory of the thesis thus moves from general to specific, and from methodology to application, demonstrating both the flexibility of two-stage seamless adaptive designs and the importance of rigorous statistical underpinnings.

5.2 Innovation and Impact

This work delivers three main innovations with practical impact. First, it provides a general mean-level design framework that links interim surrogate markers to final clinical endpoints under population shift, allowing different stage-specific endpoints while preserving a common study objective; the resulting combined estimators remain unbiased under stated conditions, gain precision in common regimes, and preserve type I error through correlation-aware spending. Second, it introduces an innovative Biosimilarity Index (BI) for programmes that aim, where scientifically justified, to base biosimilarity primarily on pharmacokinetic evidence, offering a transparent go or stop statistic within a totality-of-evidence paradigm. Third, it proposes the Relative Biosimilarity Index (RBI), which extends BI by explicitly incorporating

inherent variability and providing a method to define a similarity threshold interpretable across products, thereby improving transparency and helping to standardise review and approval; this aligns with FDA's BsUFA III regulatory science focus on streamlining evidentiary packages (FDA, 2024; FDA, 2025).

The broader implications span statistical methodology, clinical trial operations, and regulatory science. In statistical methodology, the work advances two-stage seamless adaptive design by addressing the joint challenges of endpoint changes and population shifts, as developed in Chapter 2, and by proposing RBI with structured guidance for δ selection in Chapter 4; together these contributions strengthen the inferential foundations of similarity testing and expand the toolkit for adaptive designs, biosimilar trials, and related settings such as real-world evidence integration. In clinical trial operations, the Chapter 2 framework offers a principled way to preserve statistical integrity when populations evolve or when interim and final endpoints differ, which is common in long and complex trials; this has practical importance in liver disease and beyond wherever two-stage seamless designs are used. The PK-centred BI (Chapter 3) further illustrates how unnecessary duplication can be avoided in biosimilar development, reducing costs and accelerating access to affordable biologics. In regulatory science, the work responds to calls from agencies such as FDA for more efficient, scientifically grounded approaches to biosimilar evaluation: BI operationalises a PK-centred evidence package within a totality-of-evidence approach, while RBI provides a framework that can support standardised decision-making; although broader adoption by regulatory agencies will require further validation and consensus-building,

these methods provide concrete tools that can inform future guidance and review practice.

5.3 Limitations

As with any methodological work, the contributions presented here are subject to limitations.

Methodologically, the framework of Chapter 2 relies on the ability to specify reliable relationships between endpoints and populations. In practice, such relationships may be difficult to establish, particularly when data are sparse or when the mechanisms linking short term surrogate markers to long term outcomes are poorly understood. This may limit the generalisability of the approach to settings where strong prior knowledge or external validation is lacking. Similarly, the RBI framework introduced in Chapter 4 depends on stable estimation of variability in the reference group. When reference samples are small, as is often the case in biosimilar trials, this estimation can become unstable, reducing the reliability of inference unless supplemented with external data or resampling methods.

From a practical standpoint, the implementation of seamless adaptive designs requires considerable upfront planning, extensive simulations and sophisticated trial infrastructure. While these are increasingly feasible in large scale industry sponsored trials, they may pose challenges in smaller or resource constrained settings. Moreover, the assumption of consistency across populations between stages, as employed in the PK-clinical design, may not always hold in practice, necessitating additional methodological extensions.

From a regulatory and clinical perspective, novel indices such as the RBI and novel trial designs such as the PK–clinical seamless framework require validation before being broadly adopted. Regulators and clinicians may be cautious in interpreting results derived from new methodologies until there is sufficient empirical evidence from case studies. This underscores the importance of future work aimed at demonstrating the practical performance of these methods in real world contexts.

5.4 Future Directions

The research presented in this thesis opens several avenues for further development. One direction is the extension of the RBI framework to Bayesian settings. By incorporating prior information, Bayesian methods could provide posterior probabilities of biosimilarity, which may offer more flexible and interpretable decision rules. This would also facilitate borrowing of information from external sources, an approach that could be particularly valuable in the small sample settings typical of biosimilar trials.

Another promising direction is the integration of real-world evidence to supplement data from randomised trials. Because biosimilar studies may have limited reference arms, the ability to incorporate external or real-world data could stabilise variance estimation and improve power. Statistical methods for dynamic borrowing and transportability could be adapted to the RBI framework, allowing for more robust inference without sacrificing type I error control.

The methodological frameworks developed here can also be extended beyond two stage designs. As multi-stage, platform and master protocol trials become more

common in oncology, rare diseases and other therapeutic areas, adapting the endpoint–population framework and the RBI methodology to these designs would be possible to increase their relevance and impact.

Future work is also needed to address the increasing complexity of endpoints in modern trials. These include settings with multiple or composite endpoints, as well as the challenge of integrating short-term surrogate marker outcomes with long-term clinical outcomes within a unified inferential framework.

A final direction concerns the determination of decision thresholds. While this thesis provided a structured methodology for δ selection, future work may explore adaptive or data driven approaches in which the threshold adjusts dynamically to observed variability. This could help balance the competing demands of type I error control and statistical power more flexibly. Empirical validation through retrospective applications to completed trials and prospective collaborations with regulators will be essential for building confidence in these methods and accelerating their adoption.

5.5 Concluding Remarks

This thesis has developed statistical methodologies that strengthen the rigour and efficiency of two-stage seamless adaptive clinical trials. By addressing inference under endpoint differences and population shifts, proposing an integrated PK–clinical seamless design for biosimilars, and introducing the Relative Biosimilarity Index with structured guidance for similarity threshold selection, the research makes contributions both to the theory of two-stage seamless adaptive designs and to their application in regulatory science.

The overarching conclusion is that two stage seamless adaptive designs, when underpinned by principled statistical methods, can bridge the gap between theoretical innovation and practical implementation. The methods presented here provide a foundation for future research and offer practical guidance for the design of more adaptive, efficient and scientifically rigorous trials. Ultimately, this work highlights the critical role of statistical innovation in modern drug development and its potential to improve the process through which safe, effective and affordable therapies reach patients.

Appendix A – Unbiasedness and Minimum Variance of Combined Estimator

From Section 2.2.3, with the mean-level model and arm-specific WLS,

$$\theta_2 \equiv \bar{y}_1 + \hat{d} = a_1 \bar{y}_1 + a_2 \bar{y}_2, \quad a_1 = 1 + c_1, \quad a_2 = c_2,$$

where \bar{y}_1 and \bar{y}_2 are pre- and post-shift sample means (disjoint cohorts), and c_1, c_2 are the entries of $A = \Delta R^\top (X^\top W X)^{-1} X^\top W$.

Assumptions:

- $\mathbb{E}(\bar{y}_1) = \mu_b$, $\mathbb{E}(\bar{y}_2) = \mu_a$ and $Cov(\bar{y}_1, \bar{y}_2) = 0$.
- The WLS estimator is unbiased: $\mathbb{E}(\hat{\beta}) = \beta \implies \mathbb{E}(\hat{d}) = d = \mu_a - \mu_b$. (This holds under the correctly specified mean model and standard WLS regularity; W treated as fixed or estimated from data independent of \bar{y}_1, \bar{y}_2 .)

Define the second-moment quantities

$$V_1 = Var(\bar{y}_1), \quad V_2 = Var(\bar{y}_2), \quad V_\theta = Var(\theta_2), \quad C = Cov(\theta_2, \bar{y}_2).$$

Given $\theta_2 = a_1 \bar{y}_1 + a_2 \bar{y}_2$ with $Cov(\bar{y}_1, \bar{y}_2) = 0$,

$$V_\theta = a_1^2 V_1 + a_2^2 V_2, \quad C = Cov(\theta_2, \bar{y}_2) = a_2 V_2.$$

1) Unbiasedness

Both ingredients estimate the same mean:

$$\mathbb{E}(\theta_2) = \mathbb{E}(\bar{y}_1 + \hat{d}) = \mu_b + \mathbb{E}(\hat{d}) = \mu_b + d = \mu_a, \quad \mathbb{E}(\bar{y}_2) = \mu_a.$$

Hence for any deterministic ω ,

$$\mathbb{E}[\omega \bar{y}_2 + (1 - \omega) \theta_2] = \omega \mu_a + (1 - \omega) \mu_a = \mu_a.$$

In particular, with the optimal $\omega = \omega^*$ (derived below), $\hat{\mu}_a$ is unbiased.

Remark. If plugging in sample-based estimates for the variances entering ω^* , the estimator remains **asymptotically unbiased** (exactly unbiased if the plug-ins are independent of \bar{y}_1, \bar{y}_2 ; otherwise, unbiased up to $o_p(1)$).

2) Minimum Variance

Consider the class of linear unbiased estimators

$$\hat{\mu}(\omega) = \omega \bar{y}_2 + (1 - \omega)\theta_2, \quad \omega \in \mathbb{R}.$$

Its variance is

$$\begin{aligned} \text{Var}(\hat{\mu}(\omega)) &= \omega^2 V_2 + (1 - \omega)^2 V_\theta + 2\omega(1 - \omega)C \\ &= (V_2 + V_\theta - 2C)\omega^2 + (-2V_\theta + 2C)\omega + V_\theta, \end{aligned}$$

where $V_2 + V_\theta - 2C = \text{Var}(\bar{y}_2 - \theta_2) \geq 0$.

This is a strictly convex quadratic provided $\text{Var}(\bar{y}_2 - \theta_2) > 0$ (i.e., $\bar{y}_2 \neq \theta_2$ almost surely),

so it has a unique minimizer obtained by setting the derivative to zero:

$$\omega^* = \frac{V_\theta - C}{V_2 + V_\theta - 2C}.$$

Substituting $V_\theta = a_1^2 V_1 + a_2^2 V_2$ and $C = a_2 V_2$ gives an explicit form:

$$\omega^* = \frac{a_1^2 V_1 + (a_2^2 - a_2)V_2}{a_1^2 V_1 + (1 - a_2^2)V_2}.$$

The corresponding minimum variance is

$$\text{Var}(\hat{\mu}(\omega^*)) = \frac{V_2 V_\theta - C^2}{V_2 + V_\theta - 2C} = \frac{a_1^2 V_1 V_2}{a_1^2 V_1 + (1 - a_2^2)V_2}.$$

Special Cases

- Independence case ($C = 0$, equivalently $a_2 = 0$):

$\omega^* = \frac{V_\theta}{V_2 + V_\theta}$, which is the usual Graybill-Deal weight (weight is proportional to the other estimator's variance, i.e., to the precision of \bar{y}_2).

- If $a_2 = 1$ (i.e., $\theta_2 = \bar{y}_2 + a_1 \bar{y}_1$ reuses \bar{y}_2 with unit coefficient), then $C = V_2$ and $\omega^* = 1$: the optimal estimator is simply \bar{y}_2 .
- If $V_2 \rightarrow 0$ (very precise post-shift mean), then $\omega^* \rightarrow 1$ and $\text{Var}(\hat{\mu}(\omega^*)) \rightarrow 0$.
- If $V_1 \rightarrow 0$ (very precise pre-shift mean), then $\text{Var}(\hat{\mu}(\omega^*)) \rightarrow 0$ as well via the last expression.

Appendix B – Overlap-Specific Variance Derivations (per arm)

Notation alignment (from main text). Let $z = \log x$ and $y = a + bz + \epsilon$ with $\mathbb{E}(\epsilon|z) = 0$.

Patient-level variances:

$$\sigma_z^2 = \text{Var}(z), \quad \sigma_y^2 = \text{Var}(y) = b^2\sigma_z^2 + \sigma_\epsilon^2.$$

Stages means and sizes: $\bar{z}^{(1)}$ from $n^{(1)}$ Stage-1 patients; $\bar{y}^{(2)}$ from $m^{(2)}$ Stage-2 patients; overlap count O (number measured in both). For compactness write $n := n^{(1)}, m := m^{(2)}$ in the appendix.

Define

$$A := a + b\bar{z}^{(1)}, \quad B := \bar{y}^{(2)}, \quad \hat{\mu} := \omega A + (1 - \omega)B, \quad 0 \leq \omega \leq 1.$$

Assume independence across distinct subjects: $(z_j, \epsilon_j) \perp (z_k, \epsilon_k)$ for $j \neq k$.

A.1 Building blocks

1. Variance of stage means

$$\text{Var}(A) = b^2 \text{Var}(\bar{z}^{(1)}) = \frac{b^2 \sigma_z^2}{n}, \quad \text{Var}(B) = \text{Var}(\bar{y}^{(2)}) = \frac{\sigma_y^2}{m}.$$

2. Cross-stage covariance of means (by overlap)

Let J be the Stage-1 index set ($|J| = n$), K the Stage-2 set ($|K| = m$), and $S = J \cap K$ the overlap set ($|S| = O$). Then

$$\begin{aligned} \text{Cov}(\bar{z}^{(1)}, \bar{y}^{(2)}) &= \frac{1}{nm} \sum_{j \in J} \sum_{k \in K} \text{Cov}(z_j y_k) = \frac{1}{nm} \sum_{s \in S} \text{Cov}(z_s y_s) \\ &= \frac{O}{nm} \text{Cov}(z, y) = \frac{O}{nm} b^2 \sigma_z^2, \end{aligned}$$

because different subjects are independent and

$$\text{Cov}(z, y) = \text{Cov}(z, a + bz + \epsilon) = b \text{Var}(z) = b \sigma_z^2,$$

The $z - \epsilon$ covariance is 0 from $\mathbb{E}(\epsilon|z) = 0$. Hence

$$\text{Cov}(A, B) = b \cdot \text{Cov}(\bar{z}^{(1)}, \bar{y}^{(2)}) = \frac{O}{nm} b^2 \sigma_z^2.$$

3. Master identity (fixed ω)

$$\text{Var}(\hat{\mu}) = \omega^2 \text{Var}(A) + (1 - \omega)^2 \text{Var}(B) + 2\omega(1 - \omega) \text{Cov}(A, B).$$

A.2 Fully paired overlap ($O = m = n$)

Plug $\text{Var}(A) = \frac{b^2 \sigma_z^2}{n}$, $\text{Var}(B) = \text{Var}(\bar{y}^{(2)}) = \frac{\sigma_y^2}{m}$ into the master identity:

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \frac{1}{n} \{ \omega^2 b^2 \sigma_z^2 + (1 - \omega)^2 \sigma_y^2 + 2\omega(1 - \omega) b^2 \sigma_z^2 \} \\ &= \frac{1}{n} \{ b^2 \sigma_z^2 + (1 - \omega)^2 \sigma_y^2 \}, \end{aligned}$$

using $\sigma_y^2 = b^2 \sigma_z^2 + \sigma_\epsilon^2$ and $\omega^2 + 2\omega(1 - \omega) = 1$.

A.3 Partial overlap ($0 < O < \min\{n, m\}$)

Plug $\text{Var}(A) = \frac{b^2 \sigma_z^2}{n}$, $\text{Var}(B) = \text{Var}(\bar{y}^{(2)}) = \frac{\sigma_y^2}{m}$, $\text{Cov}(A, B) = \frac{O}{nm} b^2 \sigma_z^2$ into the master

identity:

$$\text{Var}(\hat{\mu}) = b^2 \sigma_z^2 \left(\frac{\omega^2}{n} + \frac{2\omega(1 - \omega)O}{nm} + (1 - \omega)^2 \frac{\sigma_y^2}{m} \right).$$

This reduces continuously to A.2 as $O \rightarrow n = m$ and to A.4 as $O \rightarrow 0$.

A.4 No overlap ($O = 0$)

Here $\text{Cov}(A, B) = 0$, so

$$\text{Var}(\hat{\mu}) = \frac{\omega^2 b^2 \sigma_z^2}{n} + \frac{(1 - \omega)^2 \sigma_y^2}{m}.$$

With the independence-optimal weight

$$\omega^* = \frac{\frac{\sigma_y^2}{m}}{\frac{b^2 \sigma_z^2}{n} + \frac{\sigma_y^2}{m}},$$

$$\text{Var}_{\min}(\hat{\mu}) = \left(\frac{n}{b^2 \sigma_z^2} + \frac{m}{\sigma_y^2} \right)^{-1}.$$

Small-sample note. Under the independence case, if the component variances are estimated to form $\hat{\omega}$, the classic GD small sample inflation $[1 + 4\hat{\omega}(1 - \hat{\omega})\{(n - 1)^{-1} + (m - 1)^{-1}\}]$ applies.

Appendix C – Unbiasedness of the PK-Informed Estimator

Claim. $\hat{\mu} = \omega A + (1 - \omega)B$ is unbiased for $\mu := \mathbb{E}(y)$.

(i) Fixed ω

$$\mathbb{E}[A] = a + b\mathbb{E}(\bar{z}^{(1)}) = a + b\mathbb{E}(z) = \mathbb{E}(y) = \mu$$

$$\mathbb{E}[B] = \mathbb{E}(\bar{y}^{(2)}) = \mu$$

Hence $\mathbb{E}[\hat{\mu}] = \omega\mu + (1 - \omega)\mu = \mu$.

(ii) Data-dependent $\hat{\omega}$ built from (co)variance estimators.

Under multivariate normal sampling, sample means are independent of (co)variance estimators; thus $\hat{\omega} \perp (A, B)$ and

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\mathbb{E}(\hat{\mu}|\hat{\omega})] = \mathbb{E}[\hat{\omega}\mu + (1 - \hat{\omega})\mu] = \mu.$$

■

When bias can occur. If the weight uses the sample means (not only second moments), or the link is misspecified, unbiasedness need not hold.

References

- Bauer, P. & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4), 1029. <https://doi.org/10.2307/2533441>
- Berry, S. M., Broglio, K. R., Groshen, S., & Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clinical Trials*, 10(5), 720–734. <https://doi.org/10.1177/1740774513497539>
- Bretz, F., Schmidli, H., König, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal*, 48(4), 623–634. <https://doi.org/10.1002/bimj.200510232>
- Bretz, F., Koenig, F., Brannath, W., Glimm, E., & Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in medicine*, 28(8), 1181–1217. <https://doi.org/10.1002/sim.3538>
- Chongwe, G., Ali, J., Kaye, D. K., Michelo, C., & Kass, N. E. (2023). Ethics of adaptive designs for randomized controlled trials. *Ethics & Human Research*, 45(5), 2–14. <https://doi.org/10.1002/eahr.500178>
- Chow, S.-C. (2013). *Biosimilars: design and analysis of follow-on biologics*. Chapman & Hall/CRC. <https://doi.org/10.1201/b15303>
- Chow S. C. (2014). Bioavailability and bioequivalence in drug development. *Wiley interdisciplinary reviews. Computational statistics*, 6(4), 304–312. <https://doi.org/10.1002/wics.1310>
- Chow, S.-C. (2018). *Analytical similarity assessment in biosimilar product development*. Chapman & Hall/CRC. <https://doi.org/10.1201/9780203705131>
- Chow, S.-C. (2020). Complex innovative design for NASH clinical trials. *Academic Journal of Gastroenterology & Hepatology*, 2(3). <https://doi.org/10.33552/ajgh.2020.02.000537>
- Chow, S.-C. (2020). *Innovative methods for rare disease drug development*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781003049364>
- Chow, S.-C., & Chang, M. (2008). Adaptive design methods in clinical trials – a review. *Orphanet Journal of Rare Diseases*, 3(1). <https://doi.org/10.1186/1750-1172-3-11>
- Chow, S.-C., & Chang, M. (2011). *Adaptive design methods in clinical trials* (2nd ed.). Chapman & Hall/CRC. <https://doi.org/10.1201/b11505>
- Chow, S.-C., Shao, J. (2005). Inference for clinical trials with some protocol amendments. *Journal of Biopharmaceutical Statistics*, 15(4), 659–666. <https://doi.org/10.1081/BIP-200062286>
- Chow, S.-C., Lu, Q., & Tse, S.-K. (2007). Statistical analysis for two-stage seamless design with different study endpoints. *Journal of Biopharmaceutical Statistics*, 17(6), 1163–1176. <https://doi.org/10.1080/10543400701645249>

- Chow, S.-C., Lin, M. (2015). Analysis of two-stage adaptive seamless trial design. *Pharmaceutica Analytica Acta*, 6(3), 1000341. <https://doi.org/10.4172/2153-2435.1000341>
- Chow, S.-C., & Liu, J.-P. (2008). *Design and analysis of bioavailability and bioequivalence studies* (3rd ed.). Chapman & Hall/CRC. <https://doi.org/10.1201/9781420011678>
- Chow, S.-C., Tu, Y.-H. (2008). On two-stage seamless adaptive design in clinical trials. *Journal of the Formosan Medical Association*, 107(12), S52–S60. [https://doi.org/10.1016/s0929-6646\(09\)60009-7](https://doi.org/10.1016/s0929-6646(09)60009-7)
- Ciolino, J. D., Kaizer, A. M., & Bonner, L. B. (2023). Guidance on interim analysis methods in clinical trials. *Journal of Clinical and Translational Science*, 7(1), e124. <https://doi.org/10.1017/cts.2023.552>
- Davit, B. M., Chen, M. L., Conner, D. P., Haidar, S. H., Kim, S., Lee, C. H., Lionberger, R. A., Makhlof, F. T., Nwakama, P. E., Patel, D. T., Schuirmann, D. J., & Yu, L. X. (2012). Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US Food and Drug Administration. *The AAPS journal*, 14(4), 915–924. <https://doi.org/10.1208/s12248-012-9406-x>
- European Medicines Agency. (2014, November). *Guideline on the pharmacokinetic and clinical evaluation of modified release dosage forms*. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-pharmacokinetic-and-clinical-evaluation-modified-release-dosage-forms_en.pdf
- Fan, L., Zhao, J., & Li, W. (2020). The extension of 2-in-1 adaptive phase 2/3 designs and its application in oncology clinical trials. *Contemporary Clinical Trials*, 98, 106148. <https://doi.org/10.1016/j.cct.2020.106148>
- Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815. <https://doi.org/10.1080/0266476042000214501>
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2), 175–185. <http://www.jstor.org/stable/2984043>
- Filozof, C., Chow, S.-C., Dimick-Santos, L., Chen, Y.-F., Williams, R. N., Goldstein, B. J., & Sanyal, A. (2017). Clinical endpoints and adaptive clinical trials in precirrhotic nonalcoholic steatohepatitis: facilitating development approaches for an emerging epidemic. *Hepatology Communications*, 1(7), 577–585. <https://doi.org/10.1002/hep4.1079>
- Friede, T., Stallard, N., & Parsons, N. (2020). Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: methods, simulation model and their implementation in R. *Biometrical Journal*, 62(5), 1264–1283. <https://doi.org/10.1002/bimj.201900020>
- Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., & Pinheiro, J. (2006). Adaptive designs in clinical drug development—an executive summary of the PHRMA

- working group. *Journal of Biopharmaceutical Statistics*, 16(3), 275–283.
<https://doi.org/10.1080/10543400600614742>
- Gao, P., Ware, J. H., & Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, 18(6), 1184–1196.
<https://doi.org/10.1080/10543400802369053>
- Graybill, F. A., & Deal, R. B. (1959). Combining unbiased estimators. *Biometrics*, 15(4), 543–550. <https://doi.org/10.2307/2527652>
- Haidar, S. H., Makhlof, F., Schuirmann, D. J., Hyslop, T., Davit, B., Conner, D., & Yu, L. X. (2008). Evaluation of a scaling approach for the bioequivalence of highly variable drugs. *The AAPS journal*, 10(3), 450–454. <https://doi.org/10.1208/s12248-008-9053-4>
- Hatfield, I., Allison, A., Flight, L., Julious, S. A., & Dimairo, M. (2016). Adaptive designs undertaken in clinical research: a review of registered clinical trials. *Trials* 17, 150 (2016). <https://doi.org/10.1186/s13063-016-1273-9>
- Huskins, W. C., Fowler, V. G., Jr, & Evans, S. (2018). Adaptive designs for clinical trials: application to healthcare epidemiology research. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 66(7), 1140–1146.
<https://doi.org/10.1093/cid/cix907>
- Jenkins, M., Stone, A., & Jennison, C. (2010). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints†. *Pharmaceutical Statistics*, 10(4), 347–356. <https://doi.org/10.1002/pst.472>
- Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman and Hall/CRC.
- Jennison, C., & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika*, 93(1), 1-21. <https://doi.org/10.1093/biomet/93.1.1>
- Jiang, L., & Yuan, Y. (2023). Seamless phase II/III design: a useful strategy to reduce the sample size for dose optimization. *Journal of the National Cancer Institute*, 115(9), 1092–1098.
<https://doi.org/10.1093/jnci/djad103>
- Jin, M., & Zhang, P. (2021). A seamless adaptive 2-in-1 design expanding a phase 2 trial for treatment or dose selection into a phase 3 trial. *Statistics in Biopharmaceutical Research*, 14(3), 334–341. <https://doi.org/10.1080/19466315.2021.1914717>
- Jin, M., & Zhang, P. (2023). An extension to a 2-in-1 adaptive design with biomarker subpopulation selection. *Contemporary Clinical Trials*, 129, 107209.
<https://doi.org/10.1016/j.cct.2023.107209>
- Kairalla, J. A., Coffey, C. S., Thomann, M. A., & Muller, K. E. (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13, 145. <https://doi.org/10.1186/1745-6215-13-145>

- Kelly, P. J., Stallard, N., & Todd, S. (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, 15(4), 641–658. <https://doi.org/10.1081/bip-200062857>
- Kunz, C. U., Friede, T., Parsons, N., Todd, S., & Stallard, N. (2014). Data-driven treatment selection for seamless phase II/III trials incorporating early-outcome data. *Pharmaceutical Statistics*, 13(4), 238–246. <https://doi.org/10.1002/pst.1619>
- Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4), 1286–1290. <https://doi.org/10.1111/j.0006-341x.1999.01286.x>
- Lu, Q., Tse, S.-K., & Chow, S.-C. (2010). Analysis of time-to-event data under a two-stage survival adaptive design in clinical trials. *Journal of Biopharmaceutical Statistics*, 20(4), 705–719. <https://doi.org/10.1080/10543401003618066>
- Lu, Q., Tse, S.-K., Chow, S.-C., & Lin, M. (2012). Analysis of time-to-event data with nonuniform patient entry and loss to follow-up under a two-stage seamless adaptive design with Weibull distribution. *Journal of Biopharmaceutical Statistics*, 22(4), 773–784. <https://doi.org/10.1080/10543406.2012.678528>
- Lu, Q., Chow, S.-C., & Tse, S.-K. (2014). On two-stage adaptive seamless design with count data from different study durations under Weibull distribution. *Drug Designing: Open Access*, 3(3), 1000114. <https://doi.org/10.4172/2169-0138.1000114>
- Ma, W., Wang, M., & Zhu, H. (2022). Seamless phase II/III clinical trials with covariate adaptive randomization. *Statistica Sinica*, 32(2), 1079-1098. <https://doi.org/10.5705/ss.202019.0354>
- Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., & Krams, M. (2006). Adaptive seamless phase II/III designs—background, operational aspects, and examples. *Drug Information Journal*, 40(4), 463–473. <https://doi.org/10.1177/216847900604000412>
- Mai, W., & Chow, S.-C. (2024). Analysis of innovative two-stage seamless adaptive design with different endpoints and population shift. *Journal of Biopharmaceutical Statistics*. Advance online publication. <https://doi.org/10.1080/10543406.2024.2330204>
- Mander, A. P., & Thompson, S. G. (2010). Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clinical Trials*, 31(6), 572–578. <https://doi.org/10.1016/j.cct.2010.07.008>
- Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9(1), 59–73. <https://doi.org/10.2307/3001633>
- Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Statistics in medicine*, 30(28), 3267–3284. <https://doi.org/10.1002/sim.4102>
- Müller, H. H., & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3), 886-891. <https://doi.org/10.1111/j.0006-341X.2001.00886.x>

- Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., & Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, *16*(1). <https://doi.org/10.1186/s12916-018-1017-7>
- Parmar, M. K. B., Barthel, F. M. S., Sydes, M., Langle, R., Kaplan, R., Eisenhauer, E., Brady, M., James, N., Bookman, M. A., Swart, A., Qian, W., & Royston, P. (2008). Speeding up the evaluation of new agents in cancer. *Journal of the National Cancer Institute*, *100*(17), 1204–1214. <https://doi.org/10.1093/jnci/djn267>
- Posch, M., & Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*, *41*(6), 689–696. [https://doi.org/10.1002/\(sici\)1521-4036\(199910\)41:6](https://doi.org/10.1002/(sici)1521-4036(199910)41:6)
- Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., & Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, *24*(24), 3697–3714. <https://doi.org/10.1002/sim.2389>
- Pritchett, Y. L., Menon, S., Marchenko, O., Antonijevic, Z., Miller, E., Sanchez-Kam, M., Morgan-Bouniol, C. C., Nguyen, H., & Prucka, W. R. (2015). Sample size re-estimation designs in confirmatory clinical trials—current state, statistical considerations, and practical guidance. *Statistics in Biopharmaceutical Research*, *7*(4), 309–321. <https://doi.org/10.1080/19466315.2015.1098564>
- Quinlan, J. A., & Krams, M. (2006). Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal*, *40*(4), 437–444. <https://doi.org/10.1177/216847900604000409>
- Robertson, D. S., Lee, K. M., López-Kolkovska, B. C., & Villar, S. S. (2023). Response-adaptive randomization in clinical trials: from myths to practical considerations. *Statistical Science*, *38*(2). <https://doi.org/10.1214/22-sts865>
- Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., & Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, *57*(3), 909–913. <https://doi.org/10.1111/j.0006-341x.2001.00909.x>
- Rosenblum, M., & Van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, *98*(4), 845–860. <https://doi.org/10.1093/biomet/asr055>
- Sampson, A. R., & Sill, M. W. (2005). Drop-the-losers design: normal case. *Biometrical Journal*, *47*(3), 257–268. <https://doi.org/10.1002/bimj.200410119>
- Schmidli, H., Bretz, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal*, *48*(4), 635–643. <https://doi.org/10.1002/bimj.200510231>
- Shao, J., & Chow, S.-C. (2002). Reproducibility probability in clinical trials. *Statistics in Medicine*, *21*(12), 1727–1742. <https://doi.org/10.1002/sim.1177>

- Simon, N., & Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4), 613–625. <https://doi.org/10.1093/biostatistics/kxt010>
- Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine*, 29(9), 959–971. <https://doi.org/10.1002/sim.3863>
- Stallard, N., Hamborg, T., Parsons, N., & Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics*, 24(1), 168–187. <https://doi.org/10.1080/10543406.2013.857238>
- Stallard, N., Kunz, C. U., Todd, S., Parsons, N., & Friede, T. (2015). Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial. *Statistics in Medicine*, 34(23), 3104–3115. <https://doi.org/10.1002/sim.6567>
- Sun, R., Cheung, S. H., & Zhang, L. -X. (2007). A generalized drop-the-loser rule for multi-treatment clinical trials. *Journal of Statistical Planning and Inference*, 137(6), 2011–2023. <https://doi.org/10.1016/j.jspi.2006.06.039>
- Sverdlov, O., & Wong, W. K. (2014). Novel statistical designs for phase I/II and phase II clinical trials with dose-finding objectives. *Therapeutic Innovation & Regulatory Science*, 48(5), 601–612. <https://doi.org/10.1177/2168479014523765>
- Takahashi, K., Ishii, R., Maruo, K., & Gosho, M. (2022). Statistical tests for two-stage adaptive seamless design using short- and long-term binary outcomes. *Statistics in Medicine*, 41(21), 4130–4142. <https://doi.org/10.1002/sim.9500>
- Teng, Z., Tian, Y., Liu, Y., & Liu, G. (2020). Seamless phase 2/3 oncology trial design with flexible sample size determination. *Statistics in Medicine*, 39(18), 2373–2386. <https://doi.org/10.1002/sim.8543>
- Thall, P. F. (2021). Adaptive enrichment designs in clinical trials. *Annual review of statistics and its application*, 8(1), 393–411. <https://doi.org/10.1146/annurev-statistics-040720-032818>
- Tothfalusi, L. & Endrenyi, L. (2016). An exact procedure for the evaluation of reference-scaled average bioequivalence. *The AAPS journal*, 18(2), 476–489. <https://doi.org/10.1208/s12248-016-9873-6>
- U.S. Food and Drug Administration. (2001, January). *Statistical approaches to establishing bioequivalence: Guidance for industry*. <https://www.fda.gov/media/70958/download>
- U.S. Food and Drug Administration. (2003, March). *Bioavailability and bioequivalence studies for orally administered drug products: General considerations* (Guidance for industry). U.S. Food and Drug Administration. https://downloads.regulations.gov/FDA-2007-D-0369-0289/attachment_39.pdf
- U.S. Food and Drug Administration. (2015, April). *Scientific considerations in demonstrating biosimilarity to a reference product* (Guidance for industry). U.S. Food and Drug Administration. <https://www.fda.gov/media/82647/download>

- U.S. Food and Drug Administration. (2018, December). *Noncirrhotic nonalcoholic steatohepatitis with liver fibrosis: Developing drugs for treatment* (Guidance for Industry). U.S. Food and Drug Administration. <https://www.fda.gov/media/119044/download>
- U.S. Food and Drug Administration. (2019, November). *Adaptive designs for clinical trials of drugs and biologics* (Guidance for Industry). U.S. Food and Drug Administration. <https://www.fda.gov/media/78495/download>
- U.S. Food and Drug Administration. (2021, August). *Bioequivalence studies with pharmacokinetic endpoints for drugs submitted under an ANDA* (Guidance for industry). U.S. Food and Drug Administration. <https://www.fda.gov/media/87219/download>
- U.S. Food and Drug Administration. (2022, September 19). *Increasing the efficiency of biosimilar development programs* [Conference session]. FDA Workshop, Silver Spring, MD, United States. <https://www.fda.gov/media/163283/download>
- U.S. Food and Drug Administration. (2022, December). *Statistical approaches to establishing bioequivalence: Draft guidance for industry*. <https://www.fda.gov/media/163638/download>
- U.S. Food and Drug Administration. (2022, December 13). *Biosimilars: overview for health care professionals*. U.S. Food and Drug Administration. <https://www.fda.gov/drugs/biosimilars/overview-health-care-professionals>
- U.S. Food and Drug Administration. (2022, December 13). *Biosimilars: review and approval*. U.S. Food and Drug Administration. <https://www.fda.gov/drugs/biosimilars/review-and-approval#data>
- U.S. Food and Drug Administration. (2024). *BsUFA III regulatory research pilot program: revised research priorities* (Research roadmap). <https://www.fda.gov/media/175799/download>
- U.S. Food and Drug Administration. (2025, January 22). *BsUFA III regulatory science pilot program: progress update*. <https://www.fda.gov/drugs/news-events-human-drugs/bsufa-iii-regulatory-science-pilot-program-progress-update-01222025>
- Van der Baan, F. H., Knol, M. J., Klungel, O. H., Egberts, A. C., Grobbee, D. E., & Roes, K. C. (2012). Potential of adaptive clinical trial designs in pharmacogenetic research. *Pharmacogenomics*, *13*(5), 571–578. <https://doi.org/10.2217/pgs.12.10>
- Wages, N. A., & Tait, C. (2015). Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *Journal of Biopharmaceutical Statistics*, *25*(5), 903–920. <https://doi.org/10.1080/10543406.2014.920873>
- Wang, S., O'Neill, R. T., & Hung, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, *6*(3), 227–244. <https://doi.org/10.1002/pst.300>

- Wang, S., Hung, H. M. J., & O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, *51*(2), 358–374. <https://doi.org/10.1002/bimj.200900003>
- Wason, J., Magirr, D., Law, M., & Jaki, T. (2012). Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, *25*(2), 716–727. <https://doi.org/10.1177/0962280212465498>
- Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Springer. <https://doi.org/10.1007/978-3-319-32562-0>
- Yada, S. (2022). Bayesian adaptive design of early-phase clinical trials for precision medicine based on cancer biomarkers. *The International Journal of Biostatistics*, *18*(1), 109–125. <https://doi.org/10.1515/ijb-2021-0009>
- Yang, J., Li, G., Yang, D., Wu, J., Wang, J., Gao, X., & Liu, P. (2024). Seamless phase 2/3 design for trials with multiple co-primary endpoints using Bayesian predictive power. *BMC Medical Research Methodology*, *24*(1). <https://doi.org/10.1186/s12874-024-02144-2>
- Zhou, X., Liu, S., Kim, E. S., Herbst, R. S., & Lee, J. J. (2008). Bayesian adaptive design for targeted therapy development in lung cancer — a step toward personalized medicine. *Clinical Trials*, *5*(3), 181–193. <https://doi.org/10.1177/1740774508091815>
- Zhu, H., & Wong, W. K. (2023). An overview of adaptive designs and some of their challenges, benefits, and innovative applications. *Journal of Medical Internet Research*, *25*, e44171. <https://doi.org/10.2196/44171>