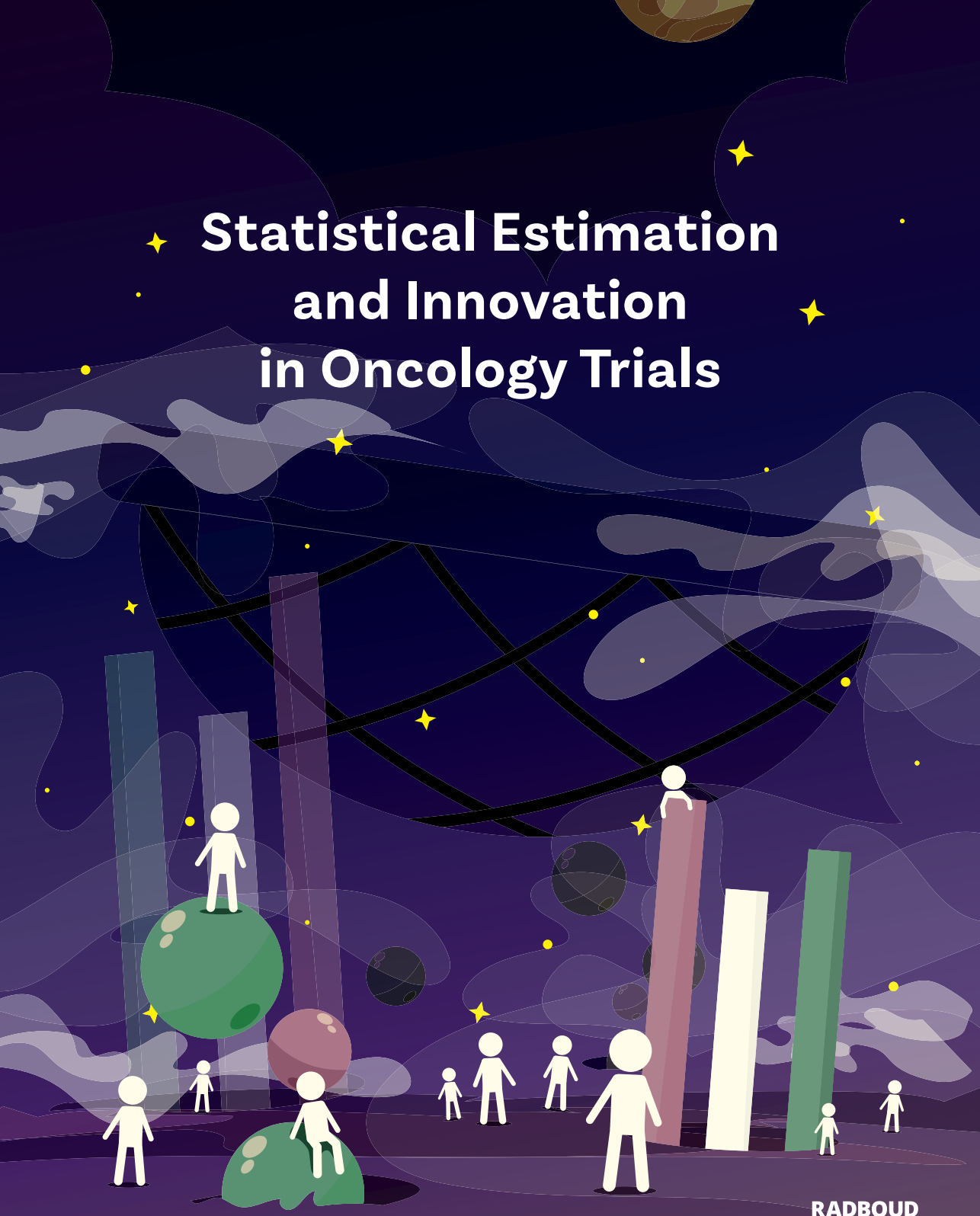
Statistical Estimation and Innovation in Oncology Trials



Antonios Daletzakis

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Statistical Estimation and Innovation in Oncology Trials

Antonios Daletzakis

Statistical Estimation and Innovation in Oncology Trials
Antonios Daletzakis

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Antonios Daletzakis
Cover: Proefschrift-aio.nl | Guus Gijben
Printing: DPN Rikken/Pumbo

ISBN: 9789465151946
DOI: 10.54195/9789465151946
Free download at: https://doi.org/10.54195/9789465151946

RADBOUD
UNIVERSITY
PRESS

Statistical Estimation and Innovation in Oncology Trials

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 8 december 2025
om 12.30 uur precies

door

Antonios Daletzakis
geboren op 24 mei 1992
te Athene (Griekenland)

Promotor

Prof. dr. C.B. Roes

Copromotoren

Dr. R.M. van den Bor (UMC Utrecht)
Dr. V. van der Noort (Nederlands Kanker Instituut)
Dr. M.A. Jonker

Manuscriptcommissie

Prof. dr. J.J. Houwing-Duistermaat
Prof. dr. A.P.H. Berkhof (Vrije Universiteit Amsterdam)
Prof. dr. L.A.L.M. Kiemeney

Statistical Estimation and Innovation in Oncology Trials

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, December 8, 2025
at 12:30 pm

by

Antonios Daletzakis
born on May 24, 1992
in Athens (Greece)

PhD supervisor

Prof. dr. C.B. Roes

PhD co-supervisors

Dr. R.M. van den Bor (UMC Utrecht)
Dr. V. van der Noort (Netherlands Cancer Institute)
Dr. M.A. Jonker

Manuscript Committee

Prof. dr. J.J. Houwing-Duistermaat
Prof. dr. A.P.H. Berkhof (Vrije Universiteit Amsterdam)
Prof. dr. L.A.L.M. Kiemeney

Contents

1

General introduction

Clinical trials have played a pivotal role in the history of medical practice, dramatically improving our ability to understand and treat diseases. From early controlled clinical trials [1], [2], and the randomized controlled trial reported in the second issue of the British Medical Journal [3], clinical research design has since continually evolved. Today, we are in the era of precision medicine, where treatments are increasingly tailored to individual patient characteristics such as genetics, lifestyle, and environmental factors [4] [5]. This evolution brings new challenges in designing and analyzing clinical trials, requiring innovative statistical methods.

Accurate estimation of treatment effects is central to clinical trial methodology, providing the quantitative evidence required for evaluating treatment efficacy, safety and cost-effectiveness. Reliable estimation informs decision-making across different stages of drug development: from initial evaluations (go/no-go decisions) to confirmatory studies, influencing regulatory approvals, reimbursement decisions and future research planning [6] , [7]. Fundamental to accurate estimation is the clear definition of the estimand, based on the precise clinical question the trial aims to address and the way the answer(s) will be quantified based on trial results.

The International Council for Harmonisation's E9(R1) [8] guideline provides a structured framework for defining estimands, clarifying precisely what clinical trials aim to estimate, typically a treatment effect, and how intercurrent events, such as treatment discontinuation or additional therapies occurring after randomization, should be handled. This guideline ensures trial objectives, data analyses, and interpretation of results are clearly aligned, enhancing the robustness and validity of conclusions.

In oncology, rapid advancements in targeted therapies and genomic diagnostics have complicated trial designs and estimation methods. Conventional randomized controlled trial designs face challenges due to the small and heterogeneous patient populations often encountered in precision oncology, particularly when studying rare cancer types or genetic aberrations across multiple tumor histologies [9]. Novel trial designs such as master protocols have emerged as innovative approaches that facilitate the simultaneous evaluation of multiple hypotheses within a single protocol structure. Master protocols allow coordinated evaluation of multiple treatments, diseases, or both, and include three distinct types: basket, umbrella, and platform trials [10]. Specifically, basket trials evaluate a single targeted therapy across different cancer types sharing a common molecular alteration. Umbrella trials

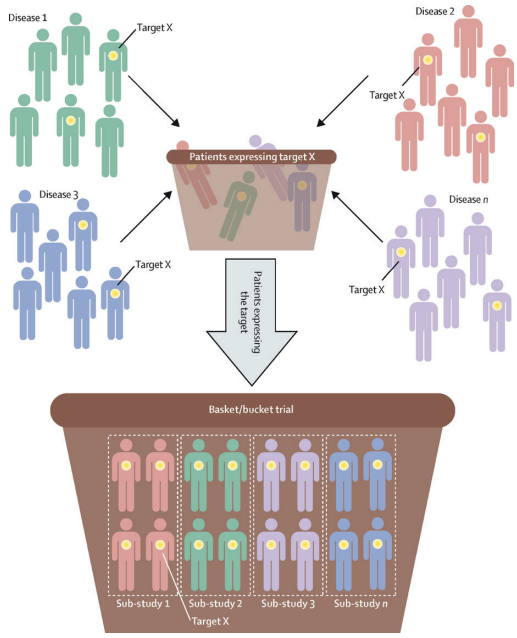


Figure 1.1: The basket trial scheme as presented in the following article https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30271-2/abstract

investigate multiple targeted therapies within a single disease, and platform trials continuously evaluate multiple treatments and disease subtypes within a perpetual framework [10]–[12].

The growing use of master protocols, such as for basket trials, needs to align with regulatory guidance provided by the ICH E9(R1) guideline [8]. Collignon et al. [13] specifically discuss the application of the estimand framework in oncology, highlighting that precise estimand definition is essential for accurately interpreting the results of basket trials and ensuring their conclusions are robust.

Basket trials (fig 1.1) offer substantial operational advantages by efficiently using resources and accelerating drug development, particularly for rare cancers or genetically-defined patient subgroups. However, these trials also introduce statistical challenges, mainly regarding how to accurately estimate treatment effectiveness across diverse patient groups. Most methods proposed in the literature for basket trials have primarily focused on binary endpoints, like tumor response rates [14]. Recently, approaches have also been developed for more complicated endpoints, such as time-to-event outcomes [15].

Basket trial designs vary significantly, including substudies with traditional single-stage trial designs to adaptive and two-stage designs, such as Simon's two-stage [16] approach (see details below) or substudies with randomized designs. A recent systematic review by Kasim et al. [17] revealed a wide variety of basket trial designs used in practice, with most adopting single-arm, phase II trials without randomization. Traditional frequentist methods for analysis [18]–[20] are most common due to their simplicity, more and more Bayesian approaches have been developed to address the heterogeneity across tumor types, including methods that allow for borrow-

ing information between substudies. Bayesian hierarchical models and model averaging methods have become increasingly common in this context [14]. If such Bayesian approaches are used to drive adaptive features for decisions in sub-studies, they will introduces a unique trial design, and careful consideration is necessary when selecting the best approach to achieve accurate and meaningful conclusions. This will be the topic of chapters 4 and 5.

Simon's two-stage design [16] is a sequential design frequently used in early-phase clinical trials, especially in oncology, to quickly evaluate whether a new treatment is promising enough for further testing. It works by dividing the trial into two stages. In the first stage, a small group of patients receives the treatment, and if the treatment shows enough promise, the study continues to the second stage, recruiting more patients. If the treatment appears ineffective after the first stage, the trial stops early to avoid wasting resources or exposing more patients to a potentially ineffective therapy. While this design is practical and efficient, it can cause estimation biases due to the early stopping rules. [21]. In STS design, the trial can only stop early for futility. The estimation problem in this case has been addressed by Jung et al. [22], who provide the uniformly minimum variance unbiased estimator (UMVUE). Several curtailment designs [23]–[25] have been proposed to allow early stopping in Simon's two-stage design, either when we can conclude that the treatment is ineffective or when it seems effective. However, these methods do not provide an UMVUE estimator when early stopping is allowed for. In Chapter 2, we provide an analytical proof of the UMVUE for trials that allow early stopping for either efficacy or futility, conditional on passing the first stage.

An endpoint in oncology trials which is becoming increasingly valuable is the Duration of Response (DoR), particularly in evaluating the efficacy of treatments that provide sustained therapeutic benefits. DoR is defined as the time from the onset of the initial response to progression of disease or death. This measure provides further insights into treatment effectiveness compared to evaluating initial response rates. Recent research highlights its increasing relevance, especially in immuno-oncology, where traditional measures like Objective Response Rate (ORR) or Progression-Free Survival (PFS) might not fully capture treatment impacts due to delayed or durable responses typical of immune checkpoint inhibitors (ICIs) (Hu et al. [26]).

DoR's sensitivity and reliability as a clinical endpoint are particularly advantageous in randomized Phase II studies. It has shown a stronger capability in identifying true positive results and correctly predicting overall survival benefits compared to conventional endpoints like ORR and PFS, thus making it highly relevant for decision-making in early-phase trials (Hu et al. [26]). Furthermore, Weber et al. [27] advocate for defining DoR within the ICH E9(R1) estimand framework, underscoring the importance of clearly specifying how intercurrent events, such as treatment discontinuation, should be handled to enhance interpretability and robustness of trial outcomes.

The practical relevance of DoR is also illustrated in frameworks such as the Dutch PASKWIL criteria [28], which incorporate DoR alongside ORR to assess clinical relevance of results from non-randomized studies in rare cancers. According to these guidelines, treatments are considered clinically relevant if they meet certain thresholds linking ORR and minimum DoR durations (e.g., ORR $> 40\%$ with DoR > 4 months) (Dutch Society for Oncology, 2021). Thus, accurately estimating DoR aligns closely with both clinical practice and regulatory frameworks.

However, estimating DoR, presents methodological challenges primarily due to right censoring and interval censoring associated with scheduling intervals for patient radiological or MRI scans used to assess tumor size and

growth. To address these challenges, Chapter 3 of this thesis will present the Restricted Mean Duration of Response (RMDoR). RMDoR represents the expected ORR and DoR within a pre-specified time window by quantifying the area between two survival curves: one for the time to progression or death and another for the time to response, progression, or death. Unlike the classical DoR, which is only defined for responders, RMDoR is a population-level measure that includes both responders and non-responders—assigning a DoR of zero to the latter. This means RMDoR naturally incorporates information about the response rate: if two treatments show similar DoR among responders but differ in overall response rates, RMDoR will favor the treatment with more responders, as it should. In contrast, classical DoR cannot capture this distinction and must therefore be interpreted together with ORR. RMDoR aims to provide a more integrated and clinically meaningful summary of treatment effects, particularly in early-phase oncology trials.

1.1 Objectives

This thesis addresses statistical and methodological challenges in estimation of treatment effects, specifically focusing on innovative trial designs. It contributes by evaluating and developing methodologies aimed at improving estimation accuracy and reliability. Specifically, the thesis objectives include:

- Developing a uniformly minimum variance unbiased estimator (UMVUE) for Simon's two-stage trials when early stopping is possible.
- Investigating the duration of response outcome, addressing possible issues and proposing a robust estimator for the RMDoR in the presence of interval censoring.
- Evaluating Bayesian estimation methods for information borrowing in basket trials. We explore different settings in Basket trials like single or two-stage designs and we provide practical guidance on parameter selection for optimal estimation performance.

1.2 Thesis outline

In **Chapter 2**, we address estimation issues associated with Simon's two-stage designs. This chapter introduces and evaluates the sample proportion estimate and we propose a uniformly minimum-variance unbiased estimator (UMVUE) specifically designed to address biases introduced when the decision of early stopping can be made either for futility or for efficacy.

In **Chapter 3**, we investigate methods for estimating the RMDoR in oncology trials. It compares various approaches through simulation studies, highlighting how effectively these methods handle interval censoring in survival data.

In **Chapter 4**, we evaluate and compare various Bayesian estimation methods that enable borrowing information across cohorts in basket trials. Simulation studies, considering single stage baskets, assess the performance of these estimators in terms of bias, mean squared error, and the extent of borrowing, providing guidance on their practical implementation.

In **Chapter 5**, we applied a selection of these Bayesian methods to the Drug Rediscovery Protocol (DRUP) study, a real-world master protocol trial in oncology. We developed a parameter optimization method based on the root mean square error measure and we evaluated the estimators performance in realistic clinical settings, offering recommendations for clinical researchers.

Through addressing these methodological challenges, this thesis aims to improve the accuracy and robustness of statistical estimation in clinical trials, supporting more informed and reliable decisions in oncology research.

2

Estimation and expected sample size in Simon's two stage designs that stop as early as possible

Authors: Antonios Daletzakis, Rutger van den Bor, Marianne A Jonker, Kit CB Roes, Harm van Tinteren

Original title: Estimation and expected sample size in Simon's two stage designs that stop as early as possible

Published in: Pharmaceutical Statistics 2022; 1-16, doi:10.1002/pst.2200

https://doi:10.1002/pst.2200

Abstract

In early phase clinical studies in oncology, Simon's two-stage designs are widely used. The trial design could be made more efficient by stopping early in the second stage when the required number of responses is reached, or when it has become clear that this target can no longer be met (a form of non-stochastic curtailment). Early stopping, however, will affect proper estimation of the response rate. We propose a uniformly minimum-variance unbiased estimator (UMVUE) for the response rate in this setting. The estimator is proven to be UMVUE using the Rao-Blackwell theorem. We evaluate the estimator's properties in terms of bias and MSE, both analytically and via simulations. We derive confidence intervals based on sample space orderings, and assess the coverage. For various design options, we evaluate the reduction in expected sample size as a function of the true response rate. Our method provides a solution for estimating response rates in case of a non-stochastic curtailment Simon's two-stage design.

2.1 Introduction

Simon's [16] two-stage (STS) design is a well known single arm study design that is commonly used in early phase oncology studies. The design is focused on deciding whether or not the response rate (p) is sufficiently promising for further evaluation. In the first stage of the trial, n_1 patients are enrolled and receive the treatment of interest. The number of subjects with a positive treatment response (x_1) is recorded. If the number of responses is less than the pre-specified threshold r_1, the trial stops. Otherwise, n_2 additional patients are included. If the total number of responses (x_t) is equal to or greater than threshold r_t, the treatment is considered promising and will typically undergo more rigorous testing.

The primary purpose of the STS design is to make a go / no-go decision whether the treatment seems promising or not. Yet, typically, it is desirable to also present an estimate of the response rate along with a confidence interval [29], [30]. Commonly, the response rate is estimated by the sample proportion. Doing so, however, ignores the sequential nature of the trial and introduces bias [21]. For this reason, various alternative estimators have been proposed. The Uniformly Minimum Variance Unbiased Estimator (UMVUE) is given by Jung et al. [22], based on Chang's et al. [31] work and they provide the proof analytically. Other estimation options (so-called 'bias-reduced' estimators) were suggested in Guo et al. [32] and Whitehead et al. [21]. Koyama et al. [29] proposed a median-unbiased estimator. In addition, estimators optimized in terms of the mean squared error (MSE) have been suggested (e.g. Kunzmann et al. [33]). If we are interested in the estimate of the response rate only when the trial at least succeeded to the second stage, further options exist. [34] [35] [36] [37]. Pocher and Desseaux [30] published an analytical overview of these methods.

These estimators, however, require the practical implementation of the study to follow the design exactly as planned. The UMVUE [22] requires that either the sample size in the first stage is fully achieved (if the trial is halted after the first stage) or the total sample size is exactly achieved (if the trial succeeds to the second stage). In practice, however, the number of responses required to make a decision may be observed before the nominal target sample size is reached. Alternatively, it may become apparent at some point during the trial that there is no possibility to actually reach the required number of responses it could be a researcher's option to stop the trial earlier than the STS design, for efficacy or futility.

This is commonly referred to as 'non-stochastic curtailment' (NSC), and the approach presented in this paper is one such approach. Note that, so-called 'stochastic curtailment' (SC) approaches have been described as well. These approaches allow for stopping early not only when it is certain that the trial will or will not meet its target, but also when one of these outcomes has become likely. Chi and Chen [24] (CC) proposed a NSC design that stops early when the decision in the original STS design can be made earlier for efficacy or futility in both stages of the trial. Ayanlowo's and Redden's[23] (AR) SC design stops early in the second stage using the conditional power calculation given a pre-specified threshold θ. Kunz and Kieser[25] (KK) proposed a SC design similar to AR allowing for early stopping using a conditional power threshold in both stages of the trial. The use of the conditional power is a methodologically interesting approach but introduces complexity in the trial design. The focus of the existing literature is about the design properties, the estimation of the observed effect is not addressed in this topic.

In this paper, we describe a simple approach to non-stochastic curtailment for the STS design (which we will refer to as the 'stopped STS' design or SSTS), and focus specifically on the estimation of the response rate by providing the UMVUE as well as a method for deriving the bounds of the confidence interval using sample space orderings. In practice, e.g. in oncology basket trials which may consists of a large number of related sub-studies with the STS design, such an approach may considerably reduce the total number of subjects that needs to be included before inference on a novel treatment across studies can be made. A simulation study was performed to compare the pro-

posed estimator properties (bias, MSE) against the sample proportion in the SSTS design and assess the coverage of the confidence interval. In addition, a comparison of the estimators is made between the standard STS and the SSTS design (using Jung's UMVUE for the STS design). Also, the expected sample size of the proposed design and the probability of early termination for efficacy and futility is compared with the standard STS design under the null hypothesis.

2.2 Proposed design SSTS

Suppose that the practical implementation of a STS design will always follow the design for the first stage of the trial, but that, in the second stage, the trial is stopped as soon as the go / no-go decision can be made. In the first stage n_1 patients enter the trial and get treatment, whereas this is at most n_2 in the second stage. The outcome of all $n_t = n_1 + n_2$ patients (responders or non-responders) who may enter the trial are assumed to be independent realizations of a $Bernoulli(p)$ distribution. The total number of responders in stage 1 is denoted as X_1 and follows a $Binomial(n_1; p)$ distribution. If the observed number of responders x_1 is less than the pre-specified threshold $x_1 < r_1$ *, then the treatment is considered to be not promising, and the trial stops for futility. Stopping for efficacy, can be claimed by the end of the first stage, if the number of responders is already larger than or equal to the threshold r_t. If $r_1 \leq x_1 < r_t$, the trial proceeds to stage 2 in which at most n_2 patients enter the study. Let Y denote the number of observed patients at time of stopping and let X_t be the number of responders at time of stopping. If it proceeds to the second stage, we could distinguish the following cases:

- **Case I**: Stop early in stage 2 due to efficacy, once the required number of responses r_t has been reached in the y^{th} patient, where $y \leq n_1 + n_2$ and $y > n_1$.
- **Case II**: Stop early in stage 2 due to futility, once there are at least $l_t = n_t - r_t + 1$ non responders since the total number of responses cannot reach r_t even if you enroll the rest of the patients and all of them would respond, thus if $x_t = y - l_t$.

In the situation where the number of responders in the first stage exceeds r_t, the trial is halted for efficacy directly after stage 1 is completed. For example, consider a trial with an optimal STS design, where $n_1 = 8$ patients are enrolled in stage 1. If the number of responders is more than equal to the threshold $r_1 = 1$, the trial proceeds to the second stage and another $n_2 = 16$ patients are added. The null hypothesis will be rejected if at least $r_t = 5$ responders are observed in total. Assume that a total of 5 patients respond after evaluating the outcome of the 12^{th} patient. Then the decision would be the same as the decision we would have made if the trial would have included all $n_t = 8 + 16$ subjects, so we could stop after the 12^{th} patient, fig 2.1. (Case I) On the other hand, if only 2 patients out of 22 respond to the treatment, then we could stop the trial early since we have only another 2 patients left and the design requires at least 3 additional responses. In that case, the treatment is not considered promising. fig 2.1. (Case II)

2.3 Uniformly Minimum Variance Unbiased Estimator (UMVUE)

The property of unbiasedness is desirable for an estimator. In Appendix A it is shown that $V = (Y, X_t)$ is complete and sufficient, and is therefore use to derive an unbiased estimator with minimum variance. The UMVUE is derived using the Rao-Blackwell theorem and the fact that sample proportion in the first stage $\widetilde{p_1} = \frac{X_1}{n_1}$ is an unbiased

*The definition of the required responses is different than in the Simon's original paper, $r_1 = r_1^{Simon's} + 1$

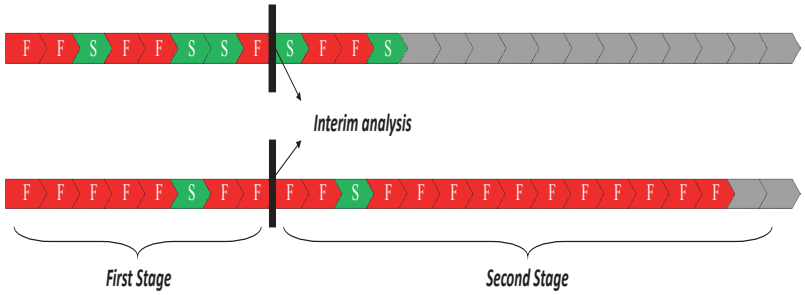


Figure 2.1: *Case I,II: Boxes with red colour are the non-responders and responders denoted with green. The black lines indicated the end of the first stage and a pre-planned interim analysis. The gray boxes indicate the unused patients.*

estimator of the response rate. Considering the sequential nature of the trial, we propose as an estimator for p the conditional expectation of the sample proportion given the sufficient statistic, $\hat{p} = E(\widetilde{p_1}|(Y, X_t))$. Following the calculations in Appendix B the UMVUE is derived:

$$\widehat{p} = \begin{cases} \frac{X_t}{n_1} & \text{if } Y = n_1 \\ \frac{\sum_{j=r_1}^{r_t-1} \binom{Y-n_1-1}{X_t-j}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1} \binom{Y-n_1-1}{X_t-j}\binom{n_1}{j}} & \text{if } Y > n_1, X_t < r_t \\ \frac{\sum_{j=r_1}^{r_t-1} \binom{Y-n_1-1}{X_t-j-1}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1} \binom{Y-n_1-1}{X_t-j-1}\binom{n_1}{j}} & \text{if } Y > n_1, X_t = r_t \end{cases} \tag{2.1}$$

where the estimator by the end of the first stage is the same if we stop due to futility, $x_t < r_1$ or efficacy $x_t \geq r_t$.

Other estimators

The UMVUE can directly be compared with the sample proportion in the stopped Simon's two-stage (SSTS) design proposed, i.e. the total number of responders divided by the total number of patients observed:

$$\hat{p}^{s_{SSTS}} = \frac{X_t}{Y}$$

In the simulations that follow, we also compare the performance of the two estimators described above against the performance of the UMVUE and sample proportion in the setting without early stopping (i.e. the standard application of the STS design), to evaluate the loss in precision that can be expected when allowing for early stopping. The sample proportion (s_{STS}) is the number of responders divided by the total number of patients: $\hat{p}^{s_{STS}} = \frac{X_t}{n_t}$ and the UMVUE [22] uses the sufficient and complete statistic (M, S), where $M \in 1, 2$ is the stage after which the trial is stopped and S is the total number of responders. If the trial stops in the first stage with $S_1 = s$, then $\hat{p}^{u_{STS}}(1, S) = \frac{S}{n_1}$. If the trial continues to the second stage with $S_2 = s$ then:

$$\hat{p}^{u_{STS}}(2, S) = \frac{\sum_{i=r_1}^{min(S,n_1)} \binom{n_1-1}{i-1}\binom{n_2-n_1}{S-1}}{\sum_{i=r_1}^{min(S,n_1)} \binom{n_1}{i}\binom{n_2-n_1}{S-1}}$$

The proposed estimator only becomes relevant if the design with early stopping is relevant and useful given study objectives and specific context. The comparison against the full design estimator is to illustrate the loss in precision such a design choice causes. †

2.3.1 Confidence interval

A common strategy in clinical trials inference is to present the response rate estimation alongside with a confidence interval. However, standard confidence intervals do not take the sequential nature of the trial into account. A natural choice would be a one-sided instead of a two-sided confidence interval for the response rate, so that it is consistent with the one-sided hypothesis testing in a STS design (remember that in a STS stage design, a one-sided hypothesis is tested to study whether the treatment is sufficiently promising). Recently Shan [38] introduced an one sided exact CI based on the p-values in two-stage designs. In the current paper a two-sided confidence interval is proposed, since the primary goal of the paper is estimation of the response rate. However, a one-sided interval can be constructed in a similar way. The two sided CI is constructed, similarly to Jung [22], by ordering the sample space based on the statistic (Y, X_t) following the SSTS design. The statistic used in this case is not a monotonic positive function. Similarly to Jennison and Turnbull [39] orderings applied in the sample space, based on the UMVUE $\widehat{p}(Y, X_t)$. The ordering is made using the sufficient statistic (Y, X_t), as illustrated in Table 4.1. Using the sample space orderings and tail probabilities (analogous to Jung [22], [39], [40]), an exact $100(1-\alpha)$ percent confidence interval for all values of p contains that satisfy both inequalities: (see in Appendix C the detailed calculations)

$$P_p(\widehat{p}(Y, X_t) \geq \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2}$$
$$P_p(\widehat{p}(Y, X_t) > \widehat{p}(y, x_t) \mid p = p_0) < 1 - \frac{a}{2}$$

A bisection method is used to calculate the lower and upperbound of the confidence interval. Based on the table 4.1 sample space orderings, the sum of the probabilities of the lower side of the CI should be at least equal to $\frac{a}{2}$ and the upper side at most equal to $1 - \frac{a}{2}$ respectively. E.g. assume that we observe $x_t = 3$ and we want to calculate a 95% confidence interval. We then sum all the probabilities below the observed design (row in bold table 4.1) given the initial value 0.25 (the initial value doesn't play any role in the algorithms convergence speed). The sum is 0.485. As the procedure continues using different values of p_L and p_U, the solution for the nominal level of a 95% confidence interval in the current example is $(0.055, 0.518)$.

2.3.2 Simulation studies

To evaluate the characteristics of the SSTS and confidence interval in comparison to the alternative estimators/settings, we performed a simulation study for various STS designs. The designs chosen were selected to provide a range of null (p_0) and alternative (p_1) hypothesis with the same type I/II error (see table 1 and 2 in supplementary material for an overview).

Data was generated as $n_1 + n_2$ Bernoulli(p) trials. For reference, the stopping rules were used following the standard application of the STS design. In this setting, the sample proportion and the UMVUE [22] are evaluated. Next, the stopping rules as outlined in section 2.2 are applied (the 'stopped STS design'), and again the sample proportion and our UMVUE estimate are determined. The estimators are evaluated in terms of the bias, MSE and the expected sample size. The performance of the confidence interval proposed in section 2.4.2 is evaluated in terms

† "u" in $\hat{p}^{u_{STS}}$ is used as abbreviation for UMVUE and "s" for Sample proportion in the formulas.

n_1	$n_{2_{obs}}$	Δ	x_t	y	$P(Y=y, X_t=x_t \mid p=0.25)$	$P(Y=y, X_t=x_t \mid p=0.055)$	$P(Y=y, X_t=x_t \mid p=0.518)$
9	-	F	0	9	0.075	0.601	0.001
9	7	F	1	16	0.030	0.212	0.000
9	8	F	2	17	0.083	0.128	0.000
9	8	T	3	17	0.028	0.008	0.001
9	7	T	3	16	0.033	0.007	0.001
9	6	T	3	15	0.040	0.007	0.002
9	5	T	3	14	0.048	0.006	0.003
9	4	T	3	13	0.056	0.006	0.006
9	**3**	**T**	**3**	**12**	**0.063**	**0.006**	**0.011**
9	2	T	3	11	0.070	0.005	0.018
9	1	T	3	10	0.075	0.004	0.030
9	-	T	3	9	0.234	0.010	0.147
9	-	T	4	9	0.117	0.001	0.236
9	-	T	5	9	0.039	0.000	0.254
9	-	T	6	9	0.009	0.000	0.181
9	-	T	7	9	0.001	0.000	0.084
9	-	T	8	9	0.000	0.000	0.022
9	-	T	9	9	0.000	0.000	0.003

Table 2.1: All possible outcomes of a design where, $n_1 = 9, n_t = 17, r_1 = 1, r_t = 3$ to test $H_0 : p_0 = 0.05$ vs $H_1 : p_1 = 0.25$ with $\alpha = 0.05$ and $\beta = 0.2$. The "Δ" column indicates if the trial is successful or not, F=False, T=True. The row in bold indicates the design we used as an example above. When the design stops in the first stage due to efficacy or futility, we set the second column $n_{2_{obs}}$ a "-"

of the coverage. 1000000 simulations were used for all measures except the coverage, due to the computational complexity of the algorithm, we perform 10000 simulation runs (The simulations were performed using R, version 4.0.2).

2.4 Comparisons and Results

2.4.1 Estimator properties

In figure 2.2, the four estimators are compared in terms of MSE and bias. The first row shows the comparison in terms of bias across different optimal STS designs with different properties and the second row shows the comparison in terms of MSE in the same designs. The unbiased estimators in the simulation study behave as expected in theory. The red dashed line indicates the UMVUE of the STS design, as expected the bias overlaps with the respective bias of the UMVUE of the SSTS design (grey continuous line), equal to 0. In both cases the MSE bottom panel, reflects the variance of the estimators. The SSTS estimator will always give an estimate based on less data, due to the design's stopping rules. For this reason the u_{SSTS} MSE is higher than the u_{STS} MSE in fig 2.2. In the standard application of the STS design, the sample proportion has a downward bias. When early stopping is allowed, however, the bias can be in both directions (although the upward bias is less pronounced for designs such us the last one presented in fig 2.2). The MSE of the SSTS sample proportion seems to be lower than the proposed UMVUE in the respective areas where the sample proportion is biased. That means that the variance in these areas of the UMVUE is much larger. It is remarkable, given the PET graph, that the MSE of the u_{SSTS} is less than the respective MSE of the sample proportion of SSTS when the PET due to efficacy in the interim (yellow line fig 2.5) crosses with the PET due to futility (red line fig 2.5) by the end of stage I. In addition, the effect of stopping early due to efficacy by the end of the first stage can be seen clearly in figure 2.2 when in the first upper panel of the figure 2.2 the true RR is $p > 0.6$. This design requires at least $r_t = 7$ responders in total out of $n_1 = 18$ by the end of the first stage.

When the true RR of p is greater than 0.6 it is highly probable to observe at least 7 responses out of the 18 patients, so the design is highly likely to stop due to efficacy by the end of the first stage. Finally, in trials designed with a high null and alternative hypothesis, like the last design in figure 2.2, the UMVUE of the SSTS design performs better when the true p is close to the alternative hypothesis, even in comparison with the STS design. Considering this, a trade-off between bias and variance can be detected. Similar results are observed across all designs that were evaluated. (see, fig $^{2.6,2.7}$ supplementary material) It is common in phase II oncology designs that the decision of a treatment activity is mostly important. Estimation when the trial stops early seems to play also a role in the planning of the trial's design. Given the results of the figure 2.2 the MSE difference in the first design panel between SSTS and STS estimates seems to be significant when the estimate is away from the region of the shaded area. A clinical researcher might be whiling to choose a curtailed design that reduces the number of expected sample size and the compromise in the trial's effect won't be that important based on the simulations.

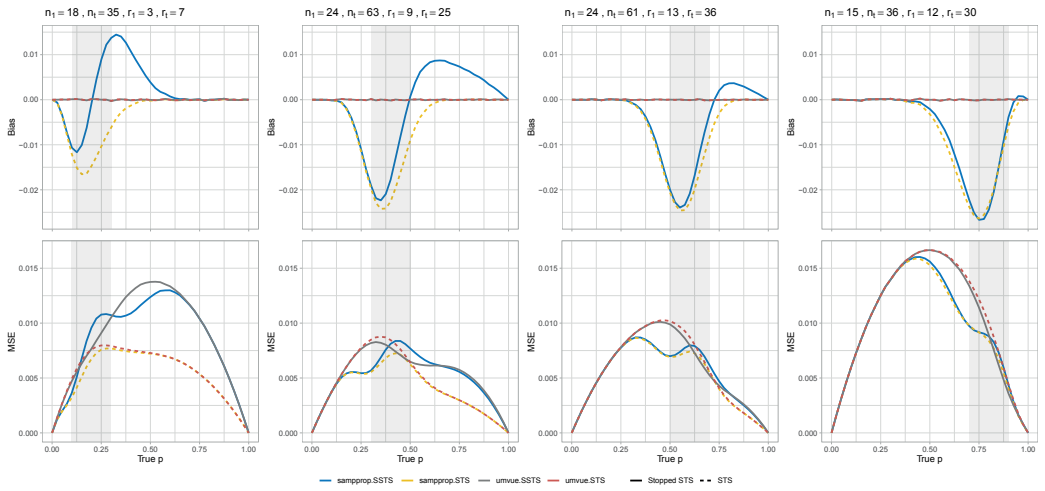


Figure 2.2: *Comparison across different null/alternative hypothesis. The blue solid line represents the sample proportion in SSTS design. The gray line denotes the UMVUE of the SSTS design. The yellow and the red dashed lines represent the sample proportion and the UMVUE of the STS respectively. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

2.4.2 Coverage probability

The coverage of the confidence interval is evaluated for a variety of designs, as discussed before. Figure 2.10 shows the simulation results. The coverage appears to be close to the nominal level of 95%, but on average higher (up to 96.8%) across all designs. When the true value of p is close to 0 or 1 a trend can be detected. The coverage is larger than the nominal level, because when all patients respond or no patient does, the method always includes the true value (0 lower bound or respective 1 upper bound). In between of the design's hypothesis range, with gray shaded colour, the coverage is close to the nominal 95% (grey shaded area). That means that the applied confidence interval is slightly conservative.

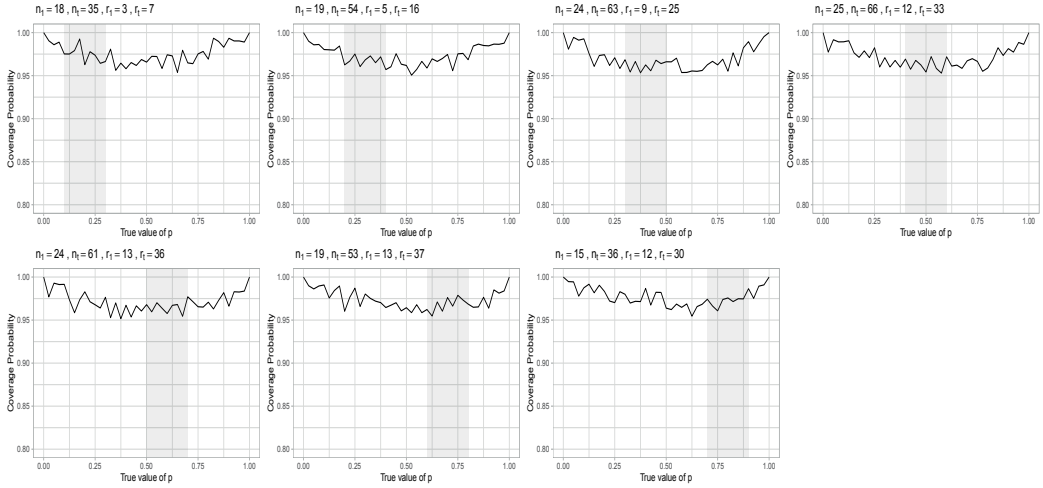


Figure 2.3: *Coverage probability across different designs of 10000 simulations. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

2.4.3 Expected sample size

The expected sample size (ESS) of SSTS is compared with the STS's ESS. Based on simulations, we provide (fig. 2.4) the whole range of true values of p, where $p \in [0, 1]$. The $ESS_{SSTS} \leq ESS_{STS}$ for any value of p. This can easily be seen in fig 2.4, the NSC design stops when the required responses can be reached or cannot be reached. An interesting point that is explained by the design properties, is that the purple line (first line, fig 2.4) becomes almost a straight line when the true value of p is greater than 0.6. The total required responses r_t in this design is a relatively small number, so the bigger the true p becomes, the more frequent responders observed and the trial stops. The ESS reduction is based on the efficacy boundary in this case. STS designs can be used in practice in a variety of ranges of p_0 and p_1 depending on the definition of "low" or "high" based on the researchers expectations of the effect. The SSTS design provides a sample size reduction for every observed outcome. In table 2.2 the expected sample size difference of STS and SSTS is depicted. Under the null hypothesis, almost 2 patients could be saved on average per trial. The SSTS would have lower expected sample size than STS, as expected theoretically. Given the simulated designs, where the power is 90%, the ESS difference with the STS design is in a range of 0.58 to 2.01 patients, with an average of 1.47. When the power is 80%, the ESS difference range is from 0.56 to 3.11, with an average of 1.56.

The ESS of the SSTS compared with other curtailed designs will have always bigger ESS. The AR design stopping rules are similar with the SSTS in the second stage, the ESS is the same when the conditional power threshold θ equals to 0, if the threshold choice is greater than 0, then the $ESS_{SSTS} \geq ESS_{AR}$. Using the same reasoning, in KK or in Chi and Chen design where early stopping is allowed in the first stage as well has a lower ESS.

2.4.4 Probability of early termination

To evaluate further the SSTS design against the STS design, the probability of early termination is calculated via simulations. The PET of STS design is coded in the fig 2.5 as the PET stage I futility (red line) and is the same for SSTS and STS design. An interesting point is to observe that the probability of early termination in stage I due to

		ESS	
p_0	p_1	SSTS	STS
$Power = 80\%$			
0.1	0.3	14.45	15.01
0.2	0.4	19.75	20.58
0.3	0.5	22.51	23.63
0.4	0.6	22.35	23.95
0.5	0.7	21.83	23.50
0.6	0.8	18.47	20.48
0.7	0.9	11.71	14.82
$Power = 90\%$			
0.1	0.3	21.94	22.53
0.2	0.4	29.16	30.43
0.3	0.5	33.41	34.72
0.4	0.6	34.48	35.98
0.5	0.7	32.31	34.01
0.6	0.8	27.59	29.47
0.7	0.9	19.22	21.23

Table 2.2: Comparison between STS optimal design and the SSTS under the null hypothesis for $p_1 - p_0 = 0.2$ at $alpha = 0.05$.

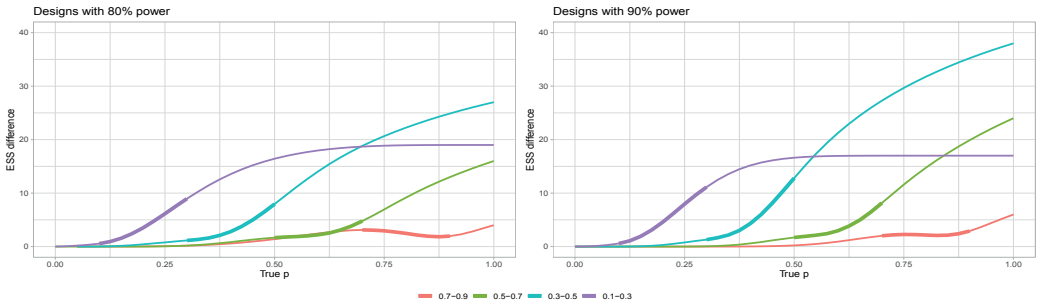


Figure 2.4: *ESS difference between STS and the SSTS across the range of possible values for p. The bold solid lines represents the design's null and alternative respective range of the studied design*

efficacy is considerably high in the upper left panel, where 7 responses are required in total to claim efficacy based on the specific design's properties. This is the only design in the graph where the total required responses is less than the first stage patients($r_t < n_1$). PET at the end of the first stage due to efficacy, introduces bias that affects the estimation. The MSE of the referred design in fig 2.2 is affected by the reduced sample size. When the yellow line of fig 2.5 starts to increase the trial stops more often by the end of the first stage.

2.4.5 Empirical application

A Malignant pleural mesothelioma study [41] is an example for Case I (fig 2.1). The trial followed the STS design ($n_1 = 18, n_2 = 15, r_1 = 5, r_t = 11$). The authors [41] estimated the response rate by means of the standard sample proportion as 47% with a 95% CI $(30\%, 65\%)$. If the UMVUE described in [22] and a CI based on Koyama [29] would have been used, the estimate would be equal to 47.5% with a 95% CI $(32.2\%, 59.7\%)$. If the researchers would have allowed for early stopping, the estimation procedure proposed in this paper would yield 47.6% with a 95% CI $(28.2\%, 70.2\%)$ and the trial could have been halted after the 22^{nd} subject out of 33 , allowing some lose of precision. The sample proportion in the SSTS would be 50% with a 95% CI $(30.7\%, 69.3\%)$)

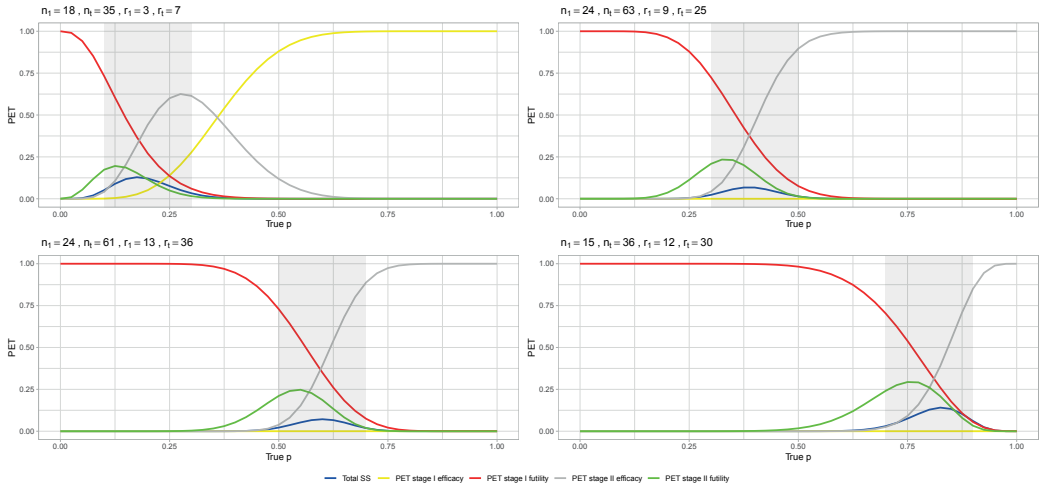


Figure 2.5: *Probability of early termination as a function of p, where the 4 different stopping rules are calculated separately. The beginning and the end of the shaded area represents p_0 and p_1 of the respective design.*

2.5 Discussion and Conclusion

In this paper, an efficient approach to the Simon's two-stage design is introduced, in situations where the go/no go decision can be made before the nominal target sample size is reached. We presented the UMVUE with corresponding confidence interval for the response rate for STS studies that stop as soon as a decision can be taken. Using simulations, the proposed estimator is compared against the sample proportion in terms of bias and MSE in various settings. In addition, it is shown how it performs against the sample proportion and UMVUE in the standard application of the STS design. Also, the coverage of the CI is investigated, as is the expected reduction in sample size. The proposed procedure is shown to be a valid approach with the potential to substantially reduce the required sample size. The SSTS design can prevent some patients from receiving ineffective treatments or, if the treatment is promising, a confirmatory trial can start sooner.

The ESS reduction could be particularly interesting in master protocols including many sub-studies that use a STS design. Recent examples like TAPUR study includes more than 336 unique trial cohorts [42] and the DRUP trial also includes more than 165 cohorts [43]. As a result, if stopped early for futility, it prevents patients exposed to treatments that are not promising. If stopped early for efficacy, rigorous evaluation starts sooner, or even treatment may become available sooner.

An important benefit of the proposed approach is the simplicity in the execution: No adjustments are needed in the design stage. Compared to most alternative curtailed designs [23],[24],[25], the SSTS design is not the most efficient in the sample size reduction, but the complexity is reduced. The stopping rules are easy to be adapted and the first stage remains the same.

Estimation becomes a part of the design, allowing for early stopping will in most cases reduce the precision of the estimate (fig 2.2). Whether this is acceptable or not is a choice that depends on the context and aims of the study, e.g. on the expected difficulty or costs of accruing new patients, as well as on the anticipated response rate and the chosen STS design. The option to provide a NSC design alongside with a UMVUE could be considered as an extra valid option when a STS design is applied. When slow accrual of patients in a rare type of cancer is observed, the

researcher should have the possibility of early stopping. In master protocols there is a variety of methods that take into account the trial's estimated effect for the homogeneity assessment of a collection of cohorts. An interesting point is that fig. 2.2 MSE graphs shows that sometimes the proposed estimator is more precise even in comparison with the complete STS design. The bias-variance trade off should be explored more, especially since we observed estimators to have a lot smaller MSE when some bias is allowed and the UMVUE are introducing substantial variability. It seems not clear which strategy is more efficient in this case. The proposed procedure is an option worth considering by any researcher involved in the design and conduct of trials following a STS design with an early stopping rule and the estimator alongside with the CI should then be used if unbiased estimation is of importance.

DATA AVAILABILITY STATEMENT

The R-script to replicate the results of the simulation study is submitted as a data file along with the manuscript.

Appendix A: Sufficiency and Completeness proof

To determine a sufficient statistic, we applied the Factorization theorem (Theorem 6.4) [44]:
Suppose that the statistical model for X consists of discrete distributions. A statistic $V = V(X)$ is sufficient if and only if there exist functions g_θ and h such that for all x and θ, $p_\theta(x) = g_\theta(V(x))h(x)$, where p_θ is the probability density of X.
In order to apply this theorem, we need to write the distribution of the observation $(X^1, \ldots, X^{n_1+n_2}, Y, \Delta)$ in the "correct" form. Where, $(X^1, \ldots, X^{n_1+n_2})$ are Bernoulli trials, with probability p that represent the responders or non-responders and $\Delta = \delta$ is a binary variable indicates the decision by the end of the trial, 0 for failure and 1 for success. The distribution of $(X^1, \ldots, X^Y, Y, \Delta)$ is equal to:

$$
\begin{aligned}
& P_p(X^1 = x^1, \ldots, X^Y = x^y, Y = y, \Delta = \delta) \\
&= P_p(X^1 = x^1, \ldots, X^y = x^y, Y = y, \Delta = \delta) \\
&= P_p(X^{n_1+1} = x^{n_1+1}, \ldots, X^y = \delta, Y = y, \Delta = \delta | X^1 = x^1, \ldots, X^{n_1} = x^{n_1}) p^{x_1}(1-p)^{n_1-x_1} \\
&= P_p(X^{n_1+1} = x^{n_1+1}, \ldots, X^y = \delta, Y = y, \Delta = \delta | X_1 = x_1) p^{x_1}(1-p)^{n_1-x_1} \\
&= P_p(X^{n_1+1} = x^{n_1+1}, \ldots, X^y = \delta | Y = y, \Delta = \delta, X_1 = x_1) P_p(Y = y, \Delta = \delta | X_1 = x_1) p^{x_1}(1-p)^{n_1-x_1} \\
&= P_p(X^{n_1+1} = x^{n_1+1}, \ldots, X^{y-1} = x^{y-1} | Y = y, \Delta = \delta, X_1 = x_1) P_p(Y = y, \Delta = \delta | X_1 = x_1) p^{x_1}(1-p)^{n_1-x_1} \\
&= h(Obs) P_p(Y = y, \Delta = \delta | X_1 = x_1) p^{x_1}(1-p)^{n_1-x_1}
\end{aligned}
\tag{2.2}
$$

with $h(Obs)$ a function of the observations only and not dependent of the unknown parameter p. This is true, because given $Y = y, \Delta = \delta$ and $X_1 = x_1$ it is given how many successes will be among $X^{n_1+1}, \ldots, X^{y-1}$. Only the order of responders and non-responders is unknown, but this does not depend on p. If $Y = n_1$ (the study is stopped after stage 1), the expression is simplified to $p^{x_1}(1-p)^{n_1-x_1}$. In the following the probability $P_p(Y = y, \Delta = \delta | X_1 = x_1) p^{x_1}(1-p)^{n_1-x_1}$ is rewritten in order to obtain the sufficient statistic with a minimum dimension.

In the following, the conditional probability $P_p(Y = y, \Delta = \delta | X_1 = x_1)$ in the expression (2.2) is expressed in the parameters. In the two cases $\delta = 1$ and $x_1 \geq r_t$, or $\delta = 0$ and $x_1 < r_1$, the study is stopped for efficacy and

futility, respectively, and:

$$P_p(Y = y, \Delta = \delta | X_1 = x_1) = \begin{cases} 1 & \text{if } y = n_1, \\ 0 & \text{otherwise.} \end{cases}$$

If $r_1 \leq x_1 < r_t$ and $\delta = 0$, the study is (early) stopped for futility in stage 2

$$\begin{aligned} P_p(Y = y, \Delta = 0 | X_1 = x_1) &= P_p(X^y = 0, X^1 + \cdots + X^{y-1} = (y-1) - (l_t - 1) | X_1 = x_1) \\ &= P_p(X^y = 0) P((X^1 + \cdots + X^{y-1}) - (X^1 + \cdots + X^{n_1}) = y - l_t - x_1 | X_1 = x_1) \\ &= (1-p)\binom{y - n_1 - 1}{y - l_t - x_1} p^{y - l_t - x_1} (1-p)^{l_t + x_1 - n_1 - 1} \\ &= \binom{y - n_1 - 1}{y - l_t - x_1} p^{y - l_t - x_1} (1-p)^{l_t + x_1 - n_1} \end{aligned}$$

If $r_1 \leq x_1 < r_t$ and $\delta = 1$ the study is (early) stopped for efficacy in stage 2

$$\begin{aligned} P_p(Y = y, \Delta = 1 | X_1 = x_1) &= P_p(X^y = 1, (X^1 + \cdots + X^{y-1}) = r_t - 1 | X_1 = x_1) \\ &= P_p(X^y = 1) P((X^1 + \cdots + X^{y-1}) - (X^1 + \cdots + X^{n_1}) = r_t - x_1 - 1) \\ &= \binom{y - n_1 - 1}{r_t - x_1 - 1} p^{r_t - x_1} (1-p)^{y + x_1 - n_1 - r_t}. \end{aligned}$$

Combining the different expression yields that $P_p(Y = y, \Delta = \delta | X_1 = x_1) p^{x_1} (1-p)^{n_1 - x_1}$ in (2.2) can be written as

$$P_p(Y = y, \Delta = 0 | X_1 = x_1) p^{x_1} (1-p)^{n_1 - x_1} = \begin{cases} \binom{y - n_1 - 1}{y - l_t - x_1} p^{y - l_t} (1-p)^{l_t} & \text{if } r_1 \leq x_1 < r_t \\ p^{x_1} (1-p)^{n_1 - x_1} & \text{if } y = n_1, x_1 < r_1 \\ 0 & \text{if } x_1 < r_1, y \neq n_1 \end{cases} \tag{2.3}$$

and

$$P_p(Y = y, \Delta = 1 | X_1 = x_1) p^{x_1} (1-p)^{n_1 - x_1} = \begin{cases} \binom{y - n_1 - 1}{r_t - x_1 - 1} p^{r_t} (1-p)^{y - r_t} & \text{if } r_1 \leq x_1 < r_t \\ p^{x_1} (1-p)^{n_1 - x_1} & \text{if } y = n_1, x_1 \geq r_t \\ 0 & \text{if } x_1 \geq r_t, y \neq n_1 \end{cases} \tag{2.4}$$

By realizing that $X_1 = X_t$ if $Y = n_1$, and $r_1 \leq X_1 < r_t$ is equivalent to $Y > n_1$ and $\Delta = 1$ is equivalent to $X_t \geq r_t$, it can be seen that the distribution $P_p(X^1 = x^1, \ldots, X^Y = x^Y, Y = y, \Delta = \delta)$ of the data can be split up in a part that depends on the observations, but is independent of the parameter p (the function h in the theorem) and a function that depends on the vector (Y, X_t) and p (the function g in the factorization theorem). The binomial coefficients in the previous displays will be part of the term that depends on the observations only. By applying the Factorization theorem it follows that the vector (Y, X_t) is sufficient.

Completeness of $V = (Y, X_t)$ can be proved via the definition of a complete statistic: A statistic V is called complete if $E_p g(V) = 0$ for all p in its parameter space, can hold only for functions g such that $P_p(g(V) = 0) = 1$ for all p. In our case, we must show that, for a function g, $h(p) = \mathrm{E}_p\, g(Y, X_t) = 0$ for all $p \in [0, 1]$, is only possible if $g(Y, X_t) = 0$ almost surely. This can be proved along the same lines as in the proof of completeness in [22].

Let g be a function with

$$h(p) \;=\; \mathrm{E}_p\, g(Y, X_t) \;=\; \sum_y \sum_{x_t} g(y, x_t) P_p(Y = y, X_t = x_t) \;=\; 0.$$

Where, the probability distribution function of V is:

$$\begin{aligned} P_p(Y = y, X_t = x_t) &= \sum_{j=0}^{n_1} P_p(X_1 = x_t, Y = y, X_t = x_t) \\ &= \begin{cases} P_p(X_1 = j, Y = y, X_t = x_t) & \text{if } y = n_1, x_t < r_1 \\ P_p(X_1 = j, Y = y, X_t = x_t) & \text{if } y = n_1, x_t \geq r_t \\ \sum_{j=r_1}^{r_t-1} P_p(Y = y, X_t = x_t | X_1 = x_t) \binom{n_1}{j} p^j (1-p)^{n_1-j} & \text{if } y > n_1, x_t < r_t \\ \sum_{j=r_1}^{r_t-1} P_p(Y = y, X_t = x_t | X_1 = j) \binom{n_1}{j} p^j (1-p)^{n_1-j} & \text{if } y > n_1, x_t = r_t \end{cases} \\ &= \begin{cases} \binom{n_1}{x_t} p^{x_t} (1-p)^{n_1-x_t} & \text{if } y = n_1, x_t < r_1 \\ \binom{n_1}{x_t} p^{x_t} (1-p)^{n_1-x_t} & \text{if } y = n_1, x_t \geq r_t \\ \sum_{j=r_1}^{r_t-1} \binom{y-n_1-1}{x_t-j} \binom{n_1}{j} p^{x_t} (1-p)^{l_t} & \text{if } y > n_1, x_t < r_t \\ \sum_{j=r_1}^{r_t-1} \binom{y-n_1-1}{r_t-j-1} \binom{n_1}{j} p^{r_t} (1-p)^{y-r_t} & \text{if } y > n_1, x_t = r_t \end{cases} \end{aligned} \tag{2.5}$$

We now need to prove that $g(y, x_t) = 0$ for all values (y, x_t) in the sample space of (Y, X_t).

By inserting the probability distribution of (Y, X_t) in the sum in the previous display, we obtain a sum of four terms. These terms correspond with the following situations: in case the trial stops in the first stage due to futility, the trial stops in the first stage due to efficacy and the last two sums refer to the second stage stopping rules for futility ($X_t = Y - l_t$) and efficacy ($X_t = r_t$), respectively. If $r_t > n_1$ early stopping for efficacy is not possible and the second term disappears.

$$\begin{aligned} h(p) &= \sum_y \sum_{x_t} g(y, x_t) P_p(Y = y, X_t = x_t) \\ &= \sum_{x_t=0}^{r_1-1} g(n_1, x_t) \binom{n_1}{x_t} p^{x_t} (1-p)^{n_1-x_t} \;+\; \sum_{x_t=r_t}^{n_1} g(n_1, x_t) \binom{n_1}{x_t} p^{x_t} (1-p)^{n_1-x_t} \\ &\quad + \sum_{y=(n_1+1)\vee l_t}^{n_t} \sum_{j=r_1}^{r_t-1} g(y, y-l_t) \binom{y-n_1-1}{y-l_t-j} \binom{n_1}{j} p^{y-l_t} (1-p)^{l_t} \\ &\quad + \sum_{y=(n_1+1)\vee r_t}^{n_t} \sum_{j=r_1}^{r_t-1} g(y, r_t) \binom{y-n_1-1}{r_t-j-1} \binom{n_1}{j} p^{r_t} (1-p)^{y-r_t}. \end{aligned}$$

For ease of notation define the functions

$$\begin{aligned} c_{1,x_t} &= g(n_1, x_t)\binom{n_1}{x_t} \\ c_{2,y} &= g(y, y-l_t)\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{y-l_t-j}\binom{n_1}{j} \\ c_{3,y} &= g(y, r_t)\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{r_t-j-1}\binom{n_1}{j}, \end{aligned}$$

all independent of p. With these definitions $h(p)$ can be written as

$$\begin{aligned} h(p) = &\sum_{x_t=0}^{r_1-1} c_{1,x_t}p^{x_t}(1-p)^{n_1-x_t} + \sum_{x_t=r_t}^{n_1} c_{1,x_t}p^{x_t}(1-p)^{n_1-x_t} \\ &+ \sum_{y=(n_1+1)\vee l_t}^{n_t} c_{2,y}p^{y-l_t}(1-p)^{l_t} + \sum_{y=(n_1+1)\vee r_t}^{n_t} c_{3,y}p^{r_t}(1-p)^{y-r_t}. \end{aligned} \tag{2.6}$$

The function $h(p)$ is written as a linear combination of terms $p^{\mu_i}(1-p)^{\nu_i}$ (where i indicates the i^{th} term in the sum of h) with $\mu_i + \nu_i$ equal to the number of patients in the trial (Y), which equals n_1, in the first and the second sum, and runs from $(n_1+1) \vee l_t$ or $(n_1+1) \vee r_t$ to n_t in the third and the fourth sums. By a close look at the terms in the sums it can be seen that for all pairs (μ_i, ν_i) and (μ_j, ν_j) with $i \neq j$ in the sums, $(\mu_i, \nu_i) \neq (\mu_j, \nu_j)$.

By taking $p = 0$, the function $h(0) = c_{1,0}$; that is for the term with $x_t = 0$. Since $h \equiv 0$ by assumption, $c_{1,0} = 0$ and therefore $g(n_1, 0) = 0$ as well (here we implicitly assume that $r_1 > 0$, otherwise the first sum in h should be left out). Now, take $p = 1$. Then $h(1) = c_{1,n_1}$ if $r_t \leq n_1$ (corresponding with the situation that is stopped after stage 1) and $h(1) = c_{3,r_t}$ if $r_t > n_1$; that is for the term with $y = r_t$ (corresponding with the situation that is continued after stage 1). Since $h \equiv 0$, it follows that $c_{1,n_1} = 0$, and thus $g(n_1, n_1) = 0$, and moreover, $c_{3,r_t} = 0$ and as a result $g(r_t, r_t) = 0$. The corresponding terms can (and should) be removed from the definition of h in (2.6).

Define $P_s(p) = h(p)/p^s$ and $Q_t(p) = h(p)/(1-p)^t$ for $p \in (0,1)$ and $s, t = 0, 1, 2, \ldots, n_t$. First, take $s = 1$. Remind that the term with p^0 equals zero (if these existed) and are left out from the definition of h. Then, $P_1(p) = h(p)/p \equiv 0$, since $h(p) \equiv 0$. Let $p \downarrow 0$. Then, $\lim_{p\downarrow 0} P_1(p) = c_{1,1}$ and $c_{1,1}$ must be equal to zero. In a similar way it can be shown that this must hold for $c_{1,2} = \ldots = c_{1,r_1-1}$; the first sum in h can be removed. Now, take $s = r_1$ and consider $P_{r_1}(p) = h(p)/p^{r_1} = 0$ and let $p \downarrow 0$, then $0 = \lim_{p\downarrow 0} P_{r_1}(p) = c_{2,l_t+r_1}$ (so that $p^{y-l_t} = p^{r_1}$). This implies that $c_{2,l_t+r_1} = 0$. In analogy, it follows that $c_{2,l_t+r_1+1} = \ldots = c_{2,l_t+r_t-1} = 0$; all constants in the third sum in h in the expression (2.6) equal zero and the third sum can be left out. By the reasoning in the previous paragraph, the first and the third sum in (2.6) equal zero.

Suppose that $r_t > n_1$, the second term disappears (it is not possible to stop after stage 1 for efficacy). The reasoning as before can be repeated with Q_t and $1-p$, in stead of P_s and p to show that $c_{3,n_1+1\vee r_t} = \ldots = c_{3,n_t} = 0$.

Now, suppose that $r_t \leq n_1$. Also now the same reasoning as before with Q_t and $1-p$, shows that $c_{1,r_t} = \ldots = c_{1,n_1} = 0$ and $c_{3,n_1+1} = \ldots = c_{3,n_t} = 0$.

Conclude that $c_{1,x_t} = 0$ for all $x_t \in \{0, \ldots, n_1\}$, $c_{2,y} = 0$ for $y \in \{(n_1+1) \vee l_t, \ldots, n_t\}$ and $c_{3,y} = 0$ for $y \in \{(n_1+1) \vee r_t, \ldots, n_t\}$. All binomials in $c_{1,x_t}, c_{2,y}$ and $c_{3,y}$ are positive, which implies that $g(Y, X_t) = 0$ almost surely.

Appendix B

As an estimator for p we take $\widehat{p} = \mathrm{E}(X_1/n_1|(Y, X_t))$, the conditional expectation of X_1/n_1 given the sufficient and complete statistic $V = (Y, X_t)$. This estimator is a function of V by construction and unbiased since: $\mathrm{E}_{(Y,X_t)}\,\mathrm{E}(X_1/n_1|(Y, X_t)) = \mathrm{E}(X_1/n_1) = p$. The explicit expression of this estimator using the sufficient statistic estimates is given by

$$\widehat{p} = \frac{1}{n_1}\mathrm{E}(X_1|Y = y, X_t = x_t) = \frac{1}{n_1}\sum_{j=0}^{n_1} jP_p(X_1 = j|Y = y, X_t = x_t)$$

$$\widehat{p} = \begin{cases} \frac{x_t}{n_1} & \text{if } y = n_1 \\ \frac{\sum_{j=0}^{r_t-1} jP_p(X_1=j,Y=y,X_t=x_t)}{n_1 P_p(Y=y,X_t=x_t)} & \text{if } y > n_1 \end{cases}$$

$$= \begin{cases} \frac{x_t}{n_1} & \text{if } y = n_1 \\ \frac{\sum_{j=r_1}^{r_t-1} j\binom{y-n_1-1}{x_t-j}\binom{n_1}{j}p^{x_t}(1-p)^{l_t}}{n_1\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{x_t-j}\binom{n_1}{j}p^{x_t}(1-p)^{l_t}} & \text{if } y > n_1, x_t < r_t \\ \frac{\sum_{j=r_1}^{r_t-1} j\binom{y-n_1-1}{r_t-j-1}\binom{n_1}{j}p^{r_t}(1-p)^{y-r_t}}{n_1\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{r_t-j-1}\binom{n_1}{j}p^{r_t}(1-p)^{y-r_t}} & \text{if } y > n_1, x_t = r_t \end{cases}$$

$$= \begin{cases} \frac{x_t}{n_1} & \text{if } y = n_1 \\ \frac{\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{x_t-j}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{x_t-j}\binom{n_1}{j}} & \text{if } y > n_1, x_t < r_t \\ \frac{\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{r_t-j-1}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1}\binom{y-n_1-1}{r_t-j-1}\binom{n_1}{j}} & \text{if } y > n_1, x_t = r_t \end{cases}$$

So the estimator for p is given by:

$$\widehat{p} = \begin{cases} \frac{X_t}{n_1} & \text{if } Y = n_1 \\ \frac{\sum_{j=r_1}^{r_t-1}\binom{Y-n_1-1}{X_t-j}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1}\binom{Y-n_1-1}{X_t-j}\binom{n_1}{j}} & \text{if } Y > n_1, X_t < r_t \\ \frac{\sum_{j=r_1}^{r_t-1}\binom{Y-n_1-1}{X_t-j-1}\binom{n_1-1}{j-1}}{\sum_{j=r_1}^{r_t-1}\binom{Y-n_1-1}{X_t-j-1}\binom{n_1}{j}} & \text{if } Y > n_1, X_t = r_t \end{cases}$$

The proposed estimator $\hat{p}$ is a function of the sufficient and complete statistic $V = (Y, X_t)$ and is unbiased. Then, by applying Theorem 6.18 in (Bijma, Jonker, van der Vaart)[44] it follows that $\hat{p}$ is a UMVUE for p. The proof of this theorem relies heavily on the Theorem of Rao-Blackwell.

Appendix C

Suppose testing the null hypothesis $H_0 : p = p_0$ against $H_1 : p \neq p_0$, the confidence interval for p equals all values for p_0 for which the test does not reject the null hypothesis. The null hypothesis is rejected if the p-value is smaller

than $\frac{\alpha}{2}$ for α the pre-specified significance level. So, if

$$P_p(\widehat{p}(Y, X_t) \geq \widehat{p}(y, x_t) \mid p = p_0) \leq \frac{a}{2}$$

OR

$$P_p(\widehat{p}(Y, X_t) \leq \widehat{p}(y, x_t) \mid p = p_0) \leq \frac{a}{2}$$

That means that the null hypothesis is not rejected if:

$$P_p(\widehat{p}(Y, X_t) \geq \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2} \tag{2.7}$$

AND

$$P_p(\widehat{p}(Y, X_t) \leq \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2} \tag{2.8}$$

So, the confidence interval for p contains all values p_0 for which both inequalities hold.
The probability $P_p(\widehat{p}(Y, X_t) \geq \widehat{p}(y, x_t) \mid p = p_0)$ is increasing as a function of p_0, so the inequality in (7) will give a lower bound of the confidence interval ($p_0 = p_L$) and similarly, the inequality in (8) will give an upper bound ($p_0 = p_U$).

Note: the latter inequality, can be rewritten as:

$$\begin{aligned} & P_p(\widehat{p}(Y, X_t) \leq \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2} \\ \iff & 1 - P_p(\widehat{p}(Y, X_t) > \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2} \\ \iff & P_p(\widehat{p}(Y, X_t) > \widehat{p}(y, x_t) \mid p = p_0) < 1 - \frac{a}{2} \end{aligned}$$

So, the confidence interval for p contains all values for p_0 that satisfy both inequalities:

$$\begin{aligned} P_p(\widehat{p}(Y, X_t) \geq \widehat{p}(y, x_t) \mid p = p_0) > \frac{a}{2} \\ P_p(\widehat{p}(Y, X_t) > \widehat{p}(y, x_t) \mid p = p_0) < 1 - \frac{a}{2} \end{aligned}$$

Appendix D

In this appendix the supplementary material of the main paper is presented. Following the same procedure as before, presenting all the designs explored in the simulation study as described in section 2.3.2. In the following table the properties of the Simon's two stage designs is presented.

		Simon's two stage optimal design properties			
p_0	p_1	r_1	n_1	r_t	n_t
$Power = 80\%$					
0.1	0.3	2	10	6	29
0.2	0.4	4	13	13	43
0.3	0.5	6	16	19	46
0.4	0.6	8	16	24	46
0.5	0.7	9	15	27	43
0.6	0.8	8	11	31	43
0.7	0.9	5	6	23	27
$Power = 90\%$					
0.1	0.3	3	18	7	35
0.2	0.4	5	19	16	54
0.3	0.5	9	24	25	63
0.4	0.6	12	25	33	66
0.5	0.7	14	24	37	61
0.6	0.8	13	19	37	53
0.7	0.9	12	15	30	36
$Different p_0 - p_1$					
0.3	0.45	14	40	41	110
0.3	0.5	9	24	25	63
0.3	0.55	6	15	17	40

Table 2.3: Optimal Simon two stage designs used for evaluation of the new proposed design

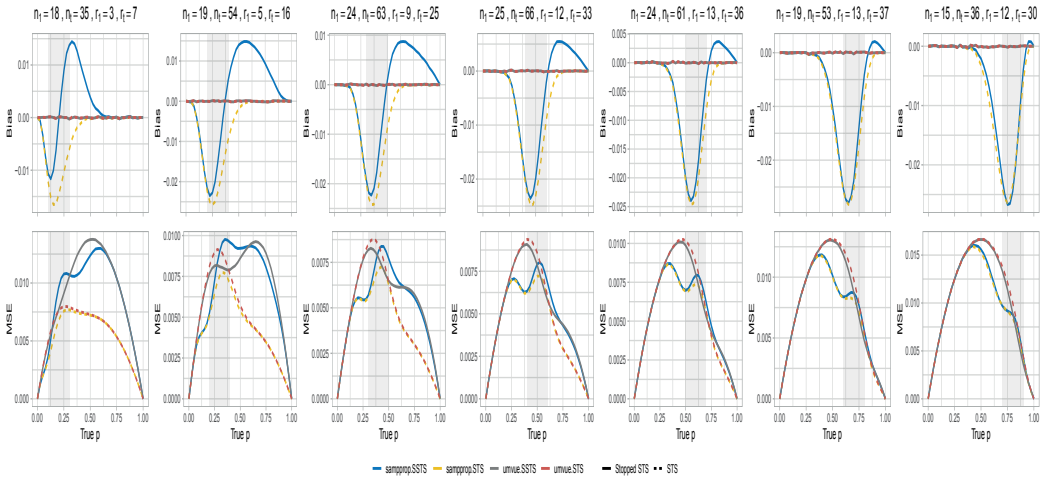


Figure 2.6: *Compare across different null/alternative hypothesis designs with 90% power, the "naive" simple proportion (yellow line) to the proposed estimator (blue line). Early stopping estimators with a continues line, full trial with a dashed line. With grey color is the sample proportion and with the red line the UMVUE. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

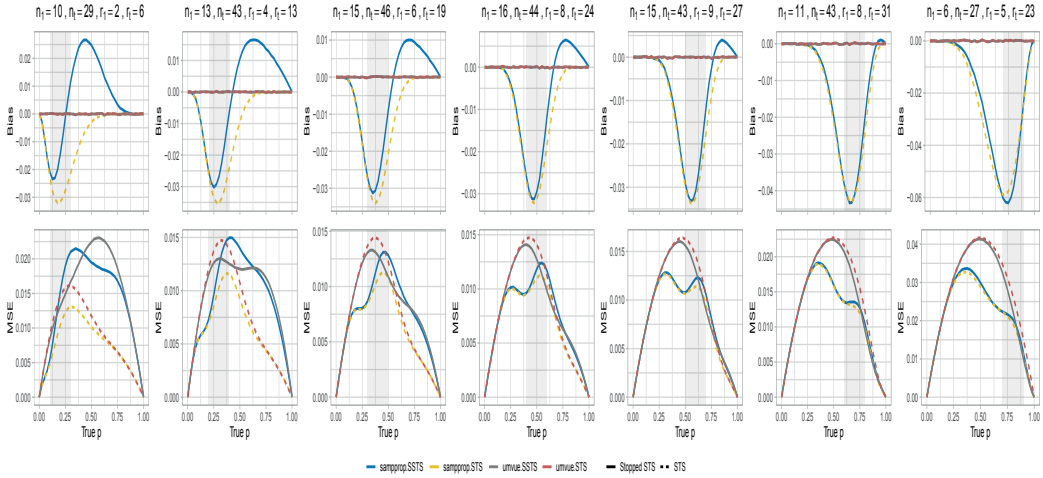


Figure 2.7: *Compare across different null/alternative hypothesis designs with 80% power, the "naive" simple proportion (yellow line) to the proposed estimator (blue line). Early stopping estimators with a continues line, full trial with a dashed line. With grey color is the sample proportion and with the red line the UMVUE. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

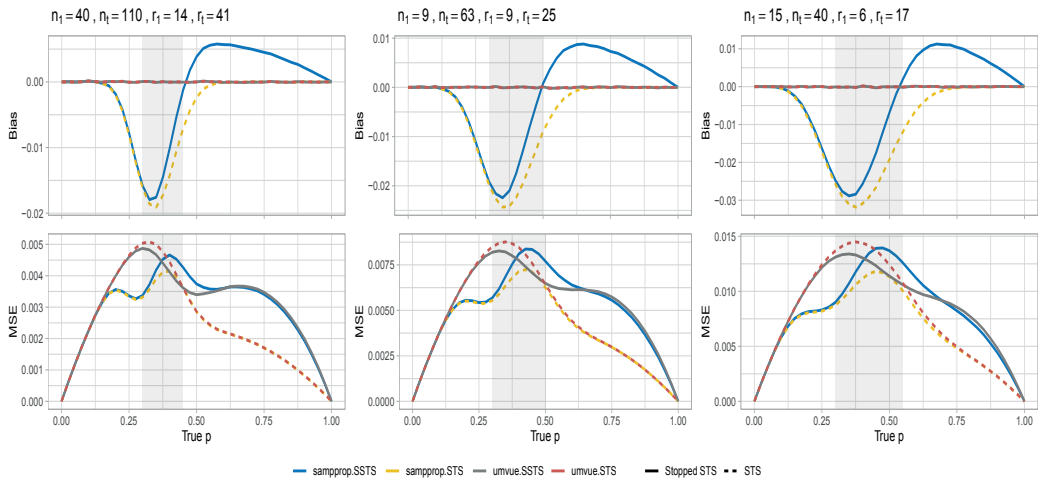


Figure 2.8: *Compare desings with different null and alternative hypothesis range, the "naive" simple proportion (yellow line) to the proposed estimator (blue line). Early stopping estimators with a continues line, full trial with a dashed line. With grey color is the sample proportion and with the red line the UMVUE. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

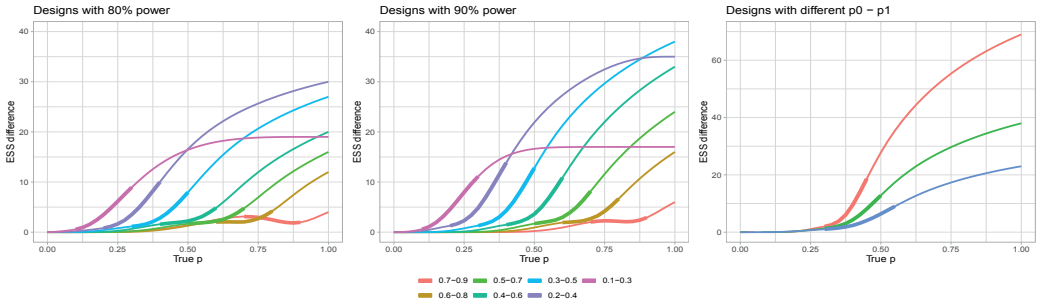


Figure 2.9: *ESS difference between Simon's two-stage design and the proposed two-stage design in every true value of p. The parts of the bold solid lines, represents the design's Null and alternative respective range of the studied design* (Note: The left graph is for 90% power and the right one for 80, in the next simulation they will be corrected in a title)

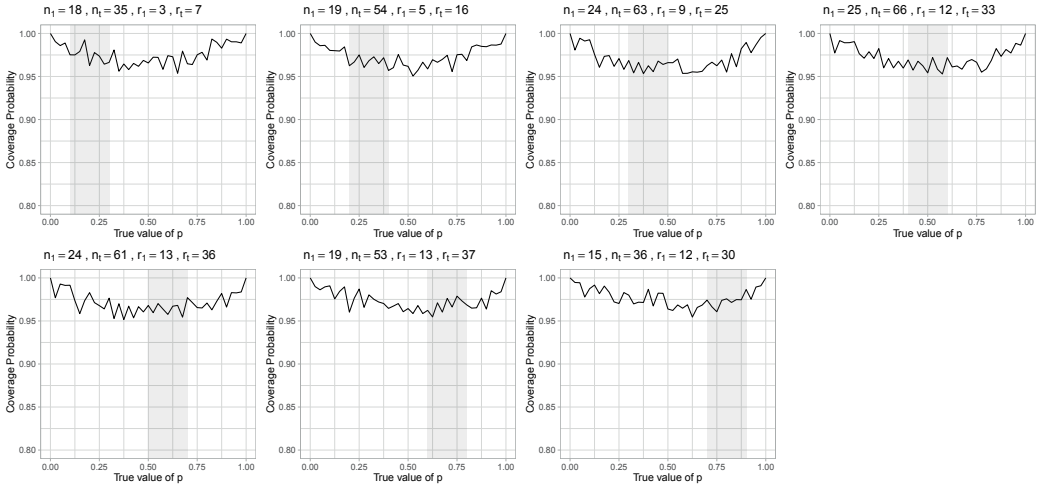


Figure 2.10: *Coverage probability across different designs of 10000 simulations. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

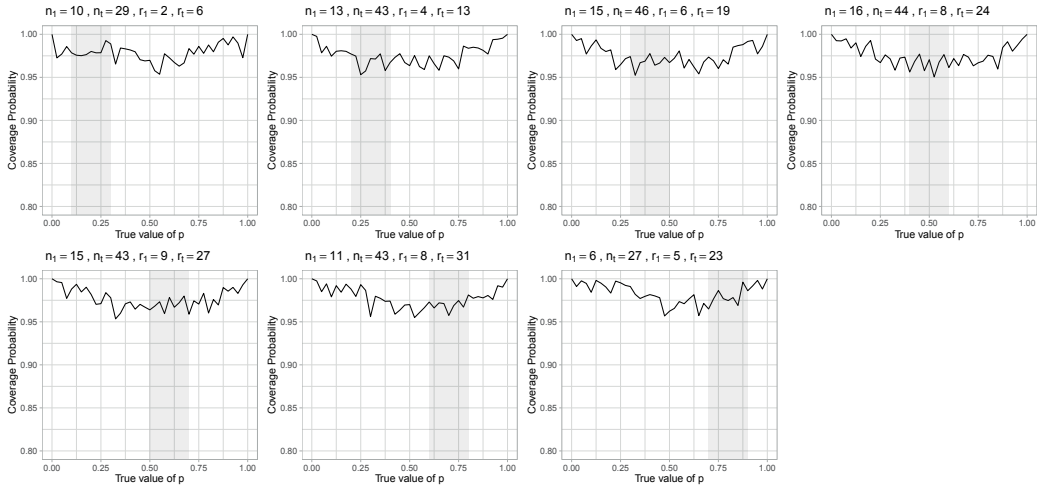


Figure 2.11: *Coverage probability across different designs of 10000 simulations. The shaded area indicates the null's and alternative's hypothesis region of the respective design.*

3

Estimation of the Restricted Mean Duration of Response (RMDoR) in oncology

Authors: Antonios Daletzakis, Kit CB Roes, Marianne A Jonker

Original title: Estimation of the Restricted Mean Duration of Response (RMDoR) in oncology

Published in: Pharmaceutical Statistics (Volume: 24, Issue: 1, Pages: e2468, Feb 2025)

https://doi.org/10.1002/pst.2468

Abstract

The Duration of Response (DoR) is defined as the time from onset of response-to-treatment up to progression-of-disease or death due to any reason, whichever occurs earlier. The expected DoR could be a suitable estimand to measure the efficacy of a treatment, but is in practice difficult to estimate, since patients' follow-up times are often right-censored. Instead, the Restricted Mean Duration of Response (RMDoR) is often used. The RMDoR in a time τ is equal to the expected DoR restricted to the interval $[0, \tau]$. In this paper we consider the behaviour of the RMDoR as a function of τ and its suitability as a measure to quantify the efficacy of a treatment. Besides, we focus on the estimation of the RMDoR. In oncology the events response-to-treatment and progression-of-disease are typically detected through time scheduled scans and are therefore interval censored. We describe multiple estimators for the RMDoR that deal with the interval censoring in different ways and study the performance of these estimators in single arm trials and randomized controlled trials.

3.1 Introduction

In oncology, but also in other research fields, the efficacy of a new treatment is preferably evaluated in a randomised clinical trial with a time-to-event endpoint. Overall survival is frequently regarded as the "gold standard". A disadvantage of this endpoint is its time-consuming nature in combination with the fact that sufficient events need to be observed to have sufficient statistical power for concluding efficacy. This has led to an exploration of surrogate clinical endpoints that are less time consuming, for instance progression free survival, objective response rate (ORR) and duration of response (DoR). An overview of different endpoints is given by Delgado.[45]

The DoR is becoming a more popular endpoint in oncology trials.[26], [46], [47] In the Netherlands, the Dutch Society for Oncology has included DoR into its criteria for clinical relevance in case of rare cancers without further treatment options, when investigated in a non-randomized design. A new treatment is deemed clinically relevant in this setting if one of the following conditions holds for the estimates: ORR > 40% and DoR > 4 months; ORR between 30% and 40% and DoR > 8 months or ORR between 20% and 30% and DoR > 12 months [28]. It is therefore highly relevant to understand definitions and estimation properties of the duration of response, including how it relates to the design (randomised or not) and assessment procedures.

An often used definition of the DoR is: the time from response initiation to disease progression or death (which one occurs earlier) in a patient who achieves complete or partial response. That means that the expected DoR is defined as the average response time within the population of patients who do respond to the treatment. So, even if the treatment is beneficial to a small proportion of the patients only, the expected DoR may be high. Because this is counter intuitive, Huang et al[48]–[50] proposed a new definition. They defined the DoR as the time from the onset of the response to disease progression or death in a patient who receives the treatment. So, they left out the condition that the patient must be a responder and, thus, changed the target population. In the new definition, patients who do not respond (before progression or death) have a DoR equal to zero. Huang and Tian[50] proved that the expected DoR equals the area between the survival curves S_{PD} and S_{RPD}, where S_{PD} is defined as the survival curve for the time to progression or death, which one occurred earlier, and S_{RPD} is the survival curve for the time to either response, progression or death, which one happened earlier.

The survival curves S_{PD} and S_{RPD} in the formula for the expected DoR can be estimated by their corresponding Kaplan-Meier curves.[51] However, most trials stop before all patients have had an event (progression or death). Hence, the data are right censored and the Kaplan-Meier curves for S_{PD} and possibly also for S_{RPD} do not reach zero. As a consequence, the expected DoR can not be estimated properly. As an alternative the restricted mean DoR (RMDoR) is considered. The RMDoR is defined as the expected DoR that is truncated at a pre-specified value τ. It turns out that the RMDoR is equal to the area between S_{PD} and S_{RPD} on the interval $[0, \tau]$. By definition, the RMDoR(τ) increases with τ and will approximate the expected DoR for τ sufficiently large. For a sensible choice of τ the RMDoR can be estimated by estimating S_{PD} and S_{RPD} by their Kaplan-Meier curves and computing the area between these curves on $[0, \tau]$. Which values for τ are sensible depends on the survival curves S_{PD} and S_{RPD} and the follow-up time of the patients.

In oncology, a response to treatment and progression of the disease are typically detected through (scheduled) scans. These scans are made when patients visit the hospital for monitoring. That means that the exact start times of response or progression are not observed, but are known to have happened before the detection time point and after the time point of the previous scan; the start times of response and progression are said to be interval censored [52]. Since the time of the event death is observed exactly, the data are partly interval censored. In practice interval censoring is often ignored when estimating survival curves; the response and progression onset are set equal to the detection times. It is well known that by doing this the Kaplan-Meier curves overestimate their corre-

sponding survival curves S_{PD} and S_{RPD} [52]. This bias can be large, depending on the scanning schedule of the patients and the survival curves S_{PD} and S_{RPD}. The RMDoR is estimated as the area between the Kaplan-Meier curves for S_{PD} and S_{RPD}. It is not clear beforehand whether the biases in the Kaplan-Meier curves also affect the performance of the estimator of the RMDoR. This is because the biases in both Kaplan-Meier curves are in the same direction (overestimation) and a part of the bias may therefore cancel out if the RMDoR is estimated as the area in between the two Kaplan-Meier curves. In this paper one of the main aims is to study the bias in the estimator of the RMDoR if the interval censoring of the observations is ignored and the observed times are used as true onset times. A simple correction for the interval censoring would be to use the midpoints of the intervals in which the response or progression occurred, in stead of the right endpoint of the interval. The corresponding Kaplan-Meier curves are still asymptotically biased, but it is expected that the bias of the corresponding RMDoR estimator will be smaller.

The aim of this paper is twofold. First, the behaviour of the RMDoR at $[0, \tau]$ as a function of τ is studied. By definition, the RMDoR increases with τ, but the question is how strong this increase is in realistic settings. The RMDoR is used in guidelines to decide upon efficacy of a treatment and the clinical relevance of treatment effects. A strong dependence on τ may make interpretation of the RMDoR difficult and its applicability for decision making debatable. The second aim of this paper is to study multiple estimators for the RMDoR. These estimators differ in the way they deal with the interval censoring. For different choices of the underlying distributions of the time to response to the treatment, progression of the disease and death, the scanning schedule of the patients and the value of τ, the performance of the estimators of the RMDoR will be studied by means of simulation studies for single and multi arm trial designs.

The paper is structured as follows. In Section 3.2 the setting in which we work and the notation is introduced. Moreover, multiple estimators for the RMDoR are defined. Thereafter, in Section 3.3, the results from the simulation study for different scenarios are described. We finish the paper with a discussion and concluding remarks in Section 3.4.

3.2 Estimation of RMDoR

In this section we introduce the notation (Subsection 3.2.1) and define multiple estimators for the RMDoR (Subsection 3.2.2). The estimators differ in the way they deal with the interval-censoring of the observations.

3.2.1 Notation and setting

For a patient the time from entering the trial (time of randomization in a randomized trial, and typically start of treatment (cycle) in a single arm trial) to response to the treatment is denoted as R, to progression of the disease as P, and to death as D. The progression free survival (PFS) time of this patient is defined as the time from entering the trial to either progression or death, which one was experienced earlier: $T_{PD} := \min\{P, D\}$. Similarly, the response-progression-free survival (RPFS) of the patient is defined as the time from entering the trial to either response, progression or death, which one was occurred earliest: $T_{RPD} := \min\{R, P, D\}$.

Following the arguments in the paper by Huang and Tian [50], the expected DoR is defined as the expected time between the response to the treatment and the progression of the disease or death, where this time is defined as zero if the patient experiences progression or death before a response to the treatment. For $x^+ := \max\{x, 0\}$, this means that the expected DoR equals

$$\mathbb{E}\,(\text{DoR}) := \mathbb{E}\big(\big(\min\{P, D\} - R\big)^+\big) = \mathbb{E}\big(\min\{P, D\} - \min\{R, P, D\}\big) = \mathbb{E}\big(T_{PD} - T_{RPD}\big)$$

It can be deduced that the expected DoR is equal to the area between the survival curves S_{PD} and S_{RPD}, where S_{PD} and S_{RPD} are the survival curves for the variables T_{PD} and T_{RPD}, respectively:

$$\begin{aligned}\mathbb{E}\,(\text{DoR}) &= \mathbb{E}\big(T_{PD} - T_{RPD}\big) \;=\; \mathbb{E}T_{PD} - \mathbb{E}T_{RPD} \\ &= -\int_0^\infty t\,\mathrm{d}S_{PD}(t) + \int_0^\infty t\,\mathrm{d}S_{RPD}(t) \;=\; \int_0^\infty S_{PD}(t)\,\mathrm{d}t - \int_0^\infty S_{RPD}(t)\,\mathrm{d}t \\ &= \int_0^\infty S_{PD}(t) - S_{RPD}(t)\,\mathrm{d}t.\end{aligned}$$

The RMDoR on the interval $[0, \tau]$, denoted as RMDoR(τ), is defined as the expected DoR truncated at timepoint τ:

$$\text{RMDoR}(\tau) := \mathbb{E}\big(\min\{T_{PD}, \tau\} - \min\{T_{RPD}, \tau\}\big) \;=\; \int_0^\tau S_{PD}(t) - S_{RPD}(t)\,\mathrm{d}t, \tag{3.1}$$

the area between the survival curves S_{PD} and S_{RPD} on the interval $[0, \tau]$. Since, by definition, $S_{PD}(t) - S_{RPD}(t) \geq 0$ for every value of t, the value of RMDoR(τ) is non-negative and non-decreasing as a function of τ. For τ increasing to infinity, RMDoR(τ) approaches the expected DoR. The behaviour of RMDoR(τ) as a function of τ depends on the underlying survival function for the time to a response of the disease, time to progression of the disease and death. In practice, every situation is different and it may be unknown beforehand how to best choose a value for τ. More discussion on the choice of the value of τ is given in the sections 3.3 and 3.4.

Suppose that the progression of the disease and the response to the treatment can only be detected by a CT scan (or another medical procedure) that is performed during one of the scheduled visits to the hospital. Define $0 = V_0, V_1, V_2, \ldots$ as the time points (counted from the moment of entering the study) of these scheduled non-random visits. When an event (response or progression) is detected at visit time V_k, the event was actually experienced in the half open interval $(V_{k-1}, V_k]$; the event is interval censored. So, the actual time to progression, P, and the time to response, R, are never observed exactly. Instead the times of the first visit after the actual progression and response times may be observed. These times are denoted as $\tilde{P}$ and $\tilde{R}$, and must, by definition, equal one of the visit times. The time of the event "death", D, is assumed to be observed exactly. Let $\tilde{T}_{PD}$ and $\tilde{T}_{RPD}$ be defined as $\tilde{T}_{PD} := \min\{\tilde{P}, D\}$ and $\tilde{T}_{RPD} := \min\{\tilde{R}, \tilde{P}, D\}$. Furthermore, define C as the independent censoring time (lost to follow-up or end of trial). Because of the nature of the data, it might happen that for a patient, $P < D$ (the true unobserved time of progression is before death), but $\tilde{P} \geq D$ (the moment the progression would have been detected is after the patient died and therefore never observed).

3.2.2 Estimators for RMDoR

The expression of the RMDoR(τ) given in (3.1) depends on the survival curves S_{PD} and S_{RPD}. After estimating these curves, the RMDoR(τ) can be estimated by the area between these estimated curves on the interval $[0, \tau]$. If the data would not be (partially) interval censored the Kaplan-Meier curves based on the true (unobserved) survival times T_{PD} and T_{RPD} (possibly censored by C) would be asymptotically pointwise (and uniformly) unbiased and the corresponding estimator for RMDoR would be asymptotically unbiased as well. However, T_{PD} and T_{RPD} are not observed, but $\tilde{T}_{PD}$ and $\tilde{T}_{RPD}$ instead (up to the censoring). Below we discuss some estimators for the survival curves.

Ignoring the fact that the data are (partially) interval censored and estimating S_{RPD} and S_{PD} by the Kaplan-Meier curves based on the observed events that equal $\min\{\tilde{T}_{PD}, C\}$ and $\min\{\tilde{T}_{RPD}, C\}$ for every patient, would give biased estimators. Since $\tilde{T}_{PD} \geq T_{PD}$ and $\tilde{T}_{RPD} \geq T_{RPD}$, these Kaplan-Meier estimators will overestimate

(pointwise) the true survival curves S_{PD} and S_{RPD}. They are shifted to the right compared to the Kaplan-Meier curves based on the true (partly) unobserved observations, due to delayed detection of response and progression. However, both estimators overestimate the true survival curves, but it is not directly clear what the effect is on the estimator for the RMDoR (defined as the area between the Kaplan-Meier curves), as both curves are biased in the same direction and part of the bias may be cancelled out. The bias in the survival curves, and probably in the RMDoR as well, depends on the schedule of the visits. This will be considered in a simulation study in Section 3.3.

To take the interval censoring of the data into account when estimating the survival curves S_{RPD} and S_{PD}, one could set the time points of progression and response equal to the mid-points of the intervals in which they were experienced, in stead of the observed right end points (the visit times). The RMDoR(τ) is estimated as the area between the newly obtained Kaplan-Meier curves on $[0, \tau]$ again. Summarized, we consider two estimators of the RMDoR(τ) which are based on the end-point and the mid-point strategy as just described:

E1. The scanning time point at which response or progression is detected is seen as the true event time-point.

E2. The response and progression time points are set equal to the midpoint of the interval in which the event was experienced and was detected at the right endpoint.

In clinical practice, medical doctors use a variety of definitions of response and progression. Sometimes a confirmation of an observed response (progression) one month later is needed. If the response is not confirmed, the patient is said not to have responded. We, therefore, also consider the end-point and midpoint estimators of the RMDoR(τ) with this alternative definition. In analogy of the estimators E1 and E2, we define the estimators based on the following definitions:

D1. The time of progression of the disease is set equal to the time-point progression was confirmed. The response to the treatment is set equal to the time-point of first detection.

D2. Similar to D1, but now the mid-points are used.

3.3 Simulation Study

In this section we describe the results of a simulation study. We consider the behaviour of the RMDoR(τ) as a function of τ (Subsection 3.3.2), the performance of the proposed estimators for the RMDoR (Subsection 3.3.3) and whether they are sensible for detecting efficacy of a new treatment in a randomized comparative setting (Subsection 3.3.4). We start with describing the assumptions for the simulation study (Subsection 3.3.1).

3.3.1 Simulation setting

We consider two settings with different time to response and progression distributions. Because we aimed for realistic distributions we chose these based on an an existing clinical trial, the Bavencio trial [53]. The Bavencio trial is a randomised two arm trial in patients with renal cell carcinoma who did not have disease progression with first-line chemotherapy. In the experimental arm the patients are treated with Avelumab and axitinib, and in the control arm the patients got Sunitinib. In total, 886 were included in the trial, of which 442 patients in the experimental arm and 444 in the control arm. In both arms every patient had a minimum follow up time of 26 months. In the experimental arm 52.5% of the patients responds to the treatment and the reported average duration of response is 9.3 months. In the control arm these numbers equal 27.3% and 5.1 months.

To base our survival models on the data from the Bavencio trial, we chose Weibull distributions for the time (from diagnosis) to response, time to progression, and time from progression to death in both arms, with parameters such that the survival functions for PFS and RPFS are similar to the Kaplan-Meier curves found in the Bavencio study. More specifically, the parameter values were chosen so that the assumed survival functions for PFS and RPFS were equal to the corresponding Kaplan-Meier curves from the Bavencio study at some selected quantiles.

Since, in the first instance, we are interested in the behavior of the RMDoR(τ) and the performance of the estimators for individual treatment arms (and not comparatively in a two-arm clinical trial), we refer to the settings in the two arms as setting A (experimental arm) and setting B (control arm). In setting A the distribution from randomization to response and progression are assumed to be Weibull(shape=1.50, scale=6.00) and Weibull(shape=0.76, scale=21.54), respectively. The distribution for the time between progression and death was taken equal to the Weibull(shape=3.00, scale=6.00). In setting B these three distributions are taken equal to the Weibull(shape=1.70, scale=10.00), Weibull(shape=0.72, scale=13.36), and Weibull(shape=3.00, scale=6.00), respectively. The corresponding survival curves S_{PD} and S_{RPD} are given in Figure 3.1, together with the RMDoR on the interval $[0, 26]$, the grey area.

Although response and progression can, in principle, happen at any time, we assume that they can only be detected during one of the moments the patients is scanned. In the simulation studies four different scanning schedules (scenarios) are considered:

S1: Every month,

S2: Every two months,

S3: Every three months,

S4: Every 1.5 months until 18 months, next every 3 months.

In the first three schedules the patients are scanned every month, every two months and every three months. These make it possible to study how the bias of the RMDoR behaves in case the scanning frequency goes down. Scanning schedule S4 was used in the Bavencio/Javelin101 trial. The moment of death is assumed to be observed exactly (if the patient died before the end of trial).

For both settings, A and B, $M = 1000$ data-sets with data of $n = 400$ patients have been simulated. Next, based on the chosen scanning schedule the observed time to response and progression are computed. For both settings and every simulation round, the RMDoR(τ) is estimated for every scanning schedule and all estimators. This gives, for every setting, every estimator and every value of τ, M estimates of the RMDoR. These M estimates are averaged and plotted. A pointwise 95% confidence interval of the estimator for RMDoR(τ) is computed as the interval from the 2.5% sample quantile to the 97.5% sample quantile.

3.3.2 Estimating the RMDoR(τ) as a function of τ in the single arm design

In this subsection we focus on the behaviour of the estimates of RMDoR(τ) as a function of τ, in the next subsection we compare the different estimates. For both settings, every scanning schedule, and in every of the M rounds, the RMDoR is estimated based on the actual event data (say the daily scan data) and by the estimators given in the previous section. See Figure 3.2 for the averages over the M rounds. From the plots we see that the estimates of RMDoR(τ) increase with τ. There is a steep increase for small values of τ, but it persists for larger values of τ. The curves would flatten when the estimates for S_{PD} and S_{RPD} get closer to each other, for instance when both Kaplan-Meier curves reach zero. The latter will only occur if the follow-up of every patient is sufficiently long. The widths

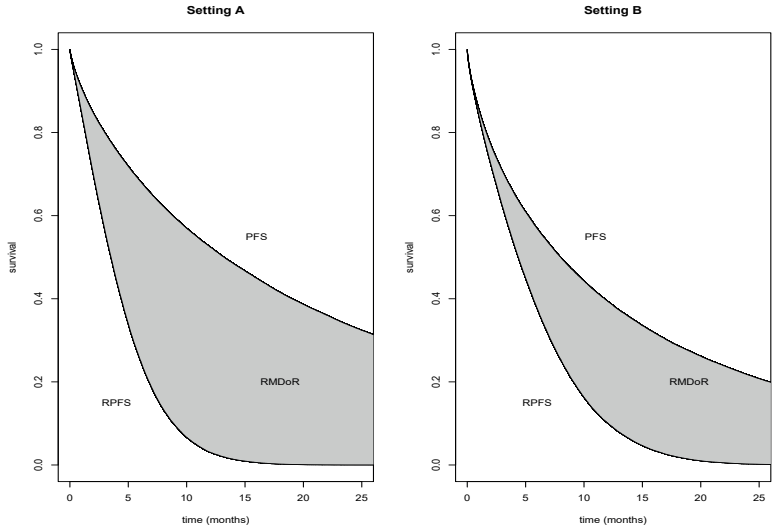


Figure 3.1: Survival curves for progression free survival (PFS) and the response and progression free survival (RPFS) in the two settings used in the simulation study. These curves are similar to the survival curves in the Bavencio trial (setting A = experimental arm, setting B = control arm). The grey area is the RMDoR on the interval $[0, 26]$.

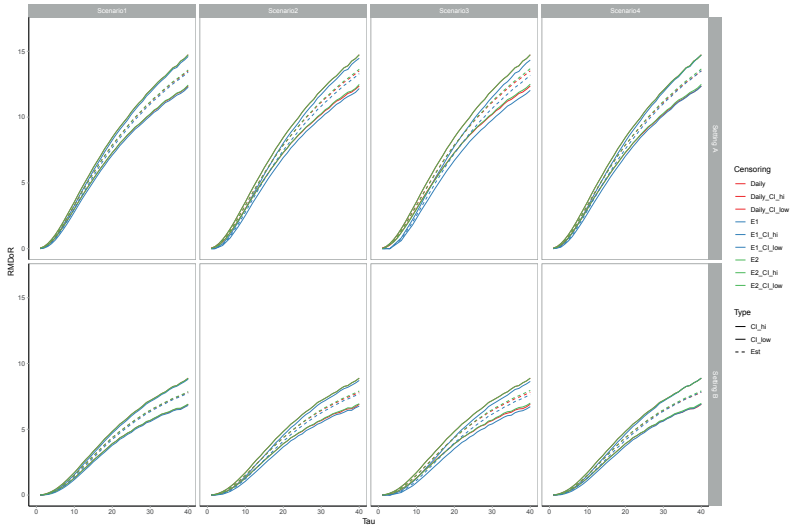


Figure 3.2: Different estimates of the RMDoR as a function of τ. Upper row: Setting A (experimental arm). Bottom row: Setting B (control arm). Every column corresponds to a scanning schedule. The estimates are represented by a dashed line, the boundaries of the confidence intervals by solid lines. Note: the estimated curves are sometimes not visible, as they overlap. The estimates D1 and D2 are not displayed, because they overlap with other curves.

of the pointwise confidence intervals increase with τ. This is due to the increasing inaccuracy of the Kaplan-Meier curves for higher values of time (and a consequence of a decreasing number of patients at risk).

3.3.3 Comparison between estimators in a single arm design

In this subsection we focus on the different estimates for the RMDoR and compare them with the estimate found based on the daily scan data in which there is no interval censoring (golden standard). The (averaged) RMDoR estimated curves were seen already in Figure 3.2. The curves are very similar; it seems that the way the interval censoring is dealt with hardly affects the estimate, especially if the times between the scans are short (schedule S1).

If we zoom in, the structural biases in the estimators are better visible. They are computed by the difference of the estimators E1, E2, D1, D2 and the one that is based on the daily scan data (the golden standard) and averaged over the M rounds (see Figure 3.3). A negative value of the difference is an underestimation of the RMDoR. The lower the curve, the more the estimator underestimates the RMDoR. Some bias-curves become positive for large values of τ, which indicates an overestimation of the RMDoR. In all plots we see the same pattern: the degree of underestimation increases with τ up to a certain value of τ (approximately 10 months), and decreases thereafter.

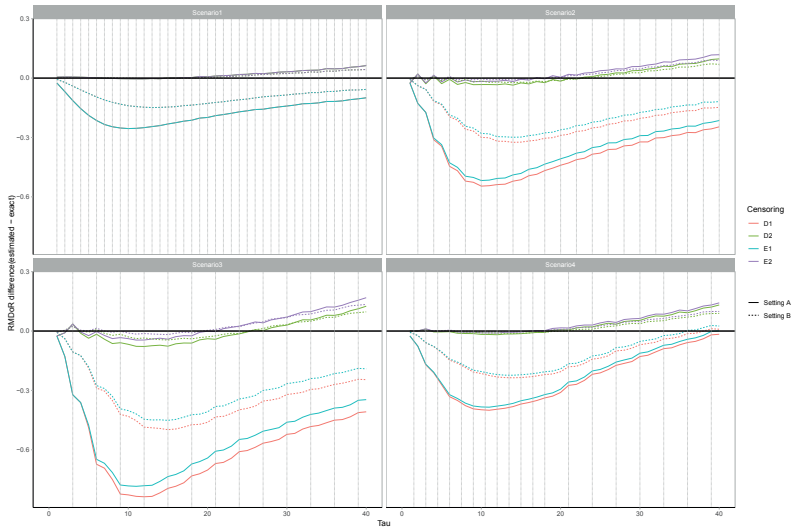


Figure 3.3: The bias (in months) in the estimation of the RMDoR(τ) for the four scanning schedules, S1,. . ., S4. The bias of the four estimators, E1, E2, D1, D2 on the interval $[0, \tau]$ are given for setting A (experimental arm) and B (control arm) (see legend) as a function of τ. The bias is computed as the estimated RMDoR based on the four estimators minus the one based on daily scans (no interval censoring). The grey vertical lines indicate the time-points the patients had a scan. In the first plot the red and green lines are not visible as they overlap with the blue and purple lines, respectively.

The scanning schedules S1, S2, and S3 (in the first three plots) are decreasing in intensity: S1 (every month), S2 (every two months), and S3 (every three months). The degree of underestimation of the RMDoR increases with the time between the scans. From the plots it can be seen that the degree of bias in the mid-point estimators (E2 and D2) is smaller than for the end-point estimators (E1 and D1). These biases are due to the way the interval censoring is taken into account when computing the Kaplan-Meier curves. If the scanning time-points are used instead of the actual time-points, the event times are assumed to be larger than they actually are ("a delayed event"). As a consequence, the jumps in the Kaplan-Meier curves are shifted to the right. Every delayed event or jump in the

Kaplan-Meier curve for S_{RPD} decreases the area between the curves and, thus, decreases the estimate of the RMDoR. On the other hand, every delayed jump in the Kaplan-Meier curve for S_{PD} increases the area between the curves and, thus, increases the estimate of the RMDoR. This is illustrated in Figure 3.4. The steeper the survival curve, the larger the overestimation of the curve and the stronger the effect on the RMDoR. In setting B (control arm) the decrease in the survival curve is more gradually in the beginning compared to setting A (experimental arm), which leads to less underestimation of the RMDoR. This is also what we see in Figure 3.3. From Figure 3.4 it is also directly clear why the bias in the mid-point estimator E2 is smaller than in the right-end point estimator.

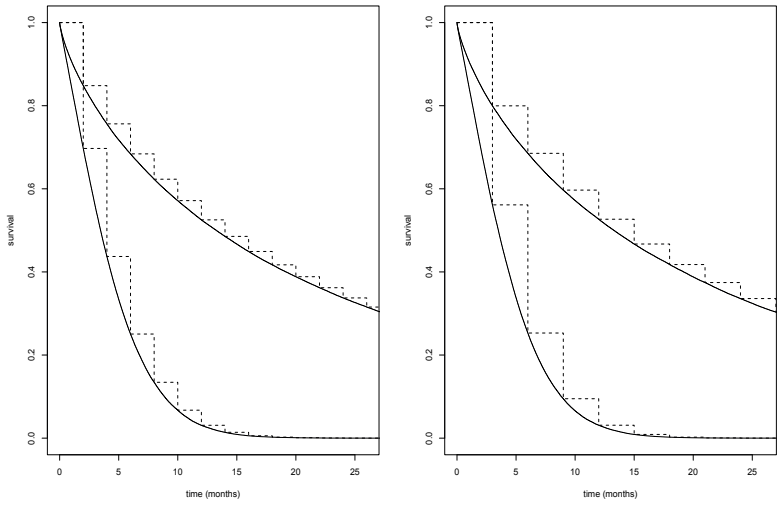


Figure 3.4: In both plots the continuous lines are the survival curves S_{RPD} and S_{RPD} in setting A. The dashed step-functions equal the Kaplan-Meier curves (for large sample size) in the case of the scanning schemes S2 (scan every two months), left plot, and S3 (scan every three months), right plot, without taking the interval censoring into account (estimator E1). For illustration purposes the event death is not taken into account.

3.3.4 Comparisons between arms in a randomized trial

In a two arm randomized trial the aim is to compare the estimated RMDoRs between the two arms for a pre-specified value of τ, by considering their difference or ratio. In this subsection we show the results of the simulation study to compare the RMDoRs in two arms (setting A and B), as a function of τ. These differences and ratios are displayed in the figures 3.5 and 3.6, respectively. For every scenario, the curves nicely overlap; we conclude that the choice of the estimator hardly affects the estimates of the differences and the ratios. Although the estimators E1 and E2 (right and mid point estimators) are biased due to the way they deal with the censoring, this bias (almost) disappears when computing the difference or ratio of the estimated RMDoRs in the two arms. It can also be seen that the estimated ratio of the RMDoRs is increasing with τ for small values of τ, but flattens out and seems to converge to fixed value close to 0.6 (for this example). In applications the value of τ is usually chosen as large as reasonable for the trial design. In this application, the ratio of the estimated RMDoRs in the two arms is constant as function of τ for large values of τ and, therefore, the ratio is the perfect quantity to compare the efficacy in the two arms. In other applications the fraction of the RMDoRs does not need to be constant for large

values of τ.

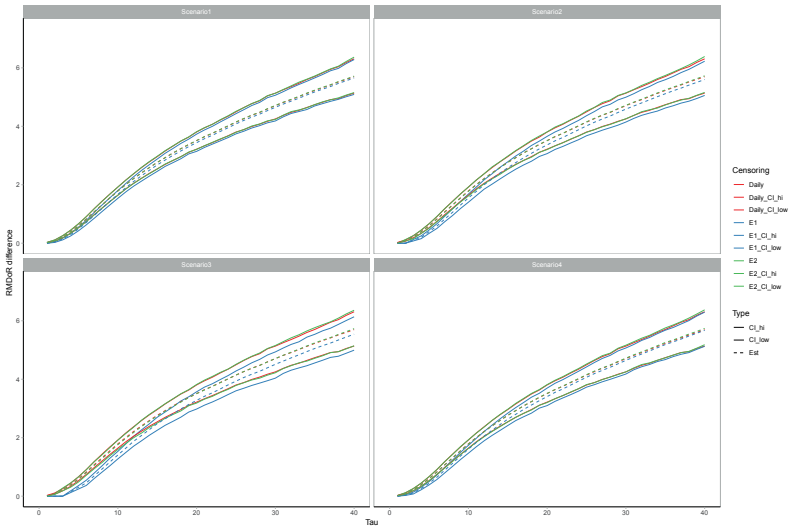


Figure 3.5: For the scanning schedules S1, ..., S4 and every estimator, the estimated difference of the RMDoR between the two arms (experimental arm minus control arm) are shown as a function of τ. Some curves are difficult to see, due to overlapping curves.

3.4 Discussion

In this paper we considered the estimation of the RMDoR. This quantity is proposed as an approximation to the expected DoR, as the latter can not be estimated accurately in the case of right-censoring. Most researchers prefer to take the follow-up time τ as large as possible, because for large values of τ the RMDoR(τ) is almost equal to the expected DoR. However, if τ is "too large", the estimate becomes unreliable, especially in trial designs with limited follow-up duration. This is because the RMDoR is estimated as the area between Kaplan-Meier curves, which become inaccurate for larger time points at which only a few patients are still at risk. In Huang and Tian [50] an algorithm is proposed for choosing the value of τ based on the data. The idea of this algorithm is to use the maximum available information from the data; so to choose the window in which the RMDoR is estimated as large as possible.

Interpretation in general of the RMDoR(τ) is difficult as it is a function of τ, and in the settings we considered, increases fast with τ. That makes an estimate of RMDoR(τ) without explicit reference to the value of τ useless. Comparing estimates of RMDoR(τ) between different patients groups or treatments is only sensible if the same value of τ is used. This will be rarely the case, unless the data of the two groups are collected within the same clinical trial. That also means that this outcome should only be used in guidelines for efficacy or clinical relevance of a treatment if a value of τ is specified. The choice of this value may be specific for the underlying disease setting. However, if more clarity on a suitable choice of the value τ is given, designs of future studies can take this into account; e.g., the follow-up of the patients should be sufficiently long to accurately estimate the RMDoR.

However, when publishing the results of a study, one need not limit oneself to an estimate of the RMDoR at a given point in time. Instead, the estimated RMDoR(τ) can be plotted as a function of τ in a figure (or table), as is done in the present paper. The time window for the plot (i.e., the largest value of τ for which the RMDoR is estimated) can

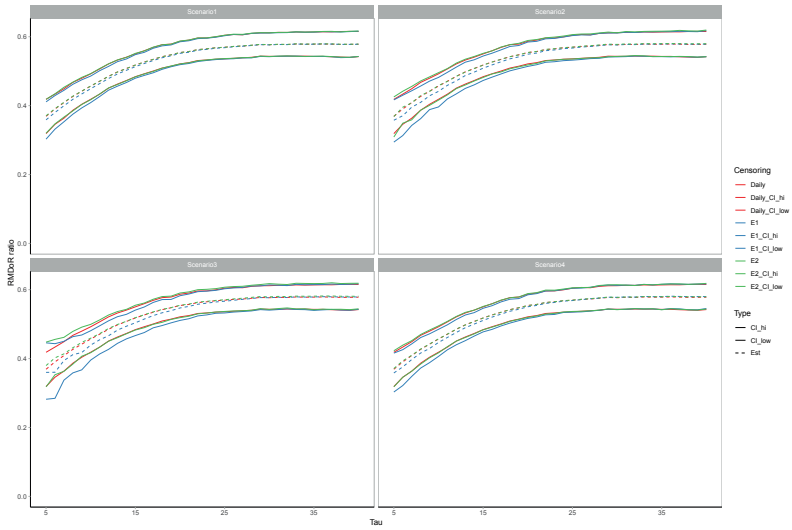


Figure 3.6: For the scanning schedules S1, ..., S4 and every estimator, the estimated ratio of the RMDoR (experimental arm/control arm) are shown as a function of τ. Some curves are difficult to see, due to overlapping of the curves.

be chosen as large as possible, for example with the data-driven algorithm proposed by Huang and Tian [50]. For comparison between studies, trial arms or patient populations, estimates of the RMDoRs at one or more specific values for τ can be extracted from the graph (or table), or comparisons can be done based on the course of the whole curve.

In the simulation studies we have seen that in two arm trials the difference of the RMDoRs in the two arms significantly increase with the value of τ, whereas the ratio was reasonable stable. This was also seen by Huang and Tian [50] in their simulations for two arm trials. So, for comparison between the RMDoRs in two arms at a chosen value of τ, the ratio of the RMDoRs is preferred due to the mild dependence of the choice of τ. However, the difference of the RMDORs in the two arms gives the absolute treatment benefit, and may therefore be more clinically meaningful than the ratio. Presenting the study results by plotting the difference and/or ratio of the RMDoRs as a function of τ may help interpretation.

In oncology the time to response and progression of the disease are often interval censored. In this paper we considered multiple ways to account for this interval censoring which led to different estimators of the RMDoR. In the simulation studies presented in the paper we have considered realistic settings, different schedules, estimators, definitions and values of τ and compared the obtained estimates of the RMDoR to the estimates that would have been found if the patients visit the medical clinic daily. In all situations and for all estimators the relative bias was small, both in the single arm and comparative setting.

Alternatively, one could try to take the interval censoring into account by using a likelihood function adjusted for the interval censoring. The NPMLEs (non-parametric maximum likelihood estimators) that are obtained by maximizing this likelihood function are not unique, but can be defined as any function that lies between the lower and upper step functions which are obtained by assuming that the events took place directly after the last visit at which the event had not taken place yet, or at the visit at which they were observed. Because the visits are the same for every patient, the number of jumps in the curves are limited and the NPMLEs will not accurately estimate S_{PD} and S_{RPD}, also not for high sample sizes.

In the paper, we described the results of multiple simulation studies to investigate the behavior of the RMDoR(τ), the performance of some estimators, and a quantity to compare the estimated RMDoRs between two arms. In the simulation study, the sample size was set at 400. This is a reasonable sample size for a phase 3 study. The same simulation study was repeated with a sample size of 100 (the figures are given in Appendix A). This sample size is more in line with the typical sample size for a phase 2 study. From the simulation study, we conclude that the results regarding bias were not dependent on sample size, but as expected the confidence intervals were slightly wider for a smaller sample size. In conclusion, from the simulation studies it follows that ignoring the interval censoring of the observations in the estimation strategy has only a minor effect on the estimate of RMDoR(τ). More important is the choice of the value τ and the associated follow-up time in clinical trial designs. If estimates of the RMDoR(τ) are used to decide upon efficacy of a treatment and clinical relevance of the effect size, more guidance is needed about the choice of τ and the trial design.

Appendix A: Results simulation results, small sample size

We have repeated the simulation study as explained in Section 3.3, but this time with a small sample size: 100 patients in each setting or arm. The results are shown in Figure 3.7 to Figure 3.10.

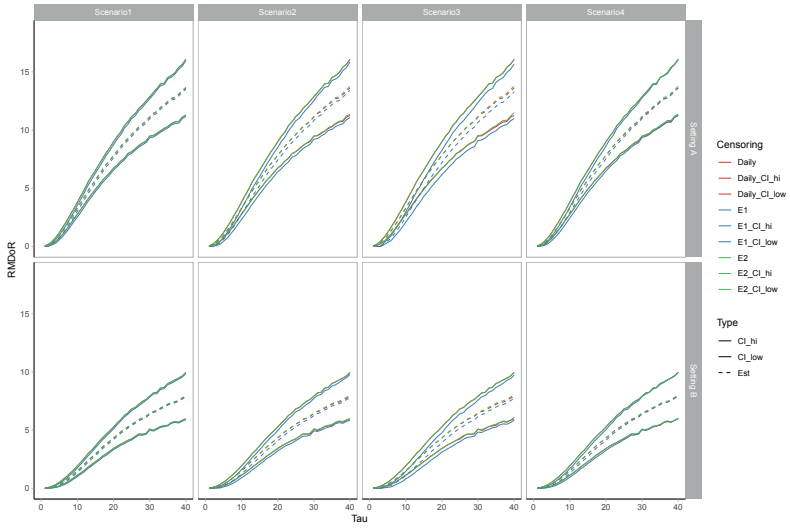


Figure 3.7: Different estimates of the RMDoR as a function of τ. Upper row: Setting A (experimental arm). Bottom row: Setting B (control arm). Every column corresponds to a scanning schedule. The estimates are represented by a dashed line, the boundaries of the confidence intervals by solid lines. Note: the estimated curves are sometimes not visible, as they overlap. The estimates D1 and D2 are not displayed, because they overlap with other curves. In this simulation 100 patients per setting is used.

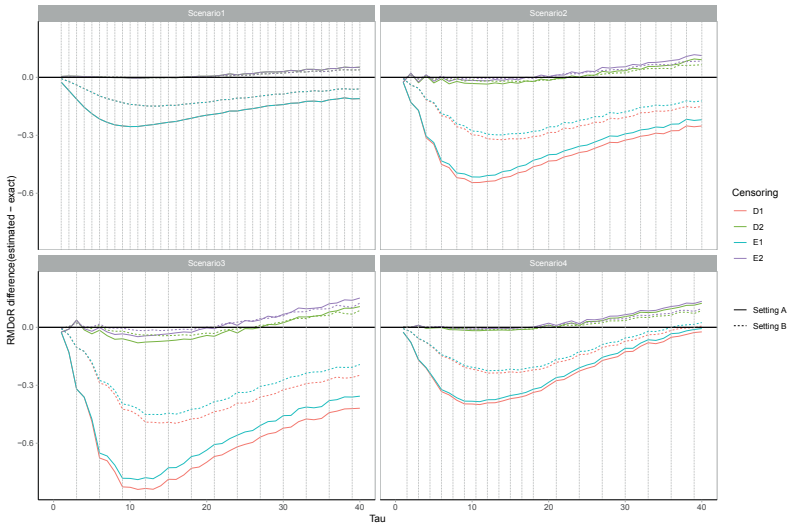


Figure 3.8: The bias (in months) in the estimation of the RMDoR(τ) for the four scanning schedules, S1,. . ., S4. The bias of the four estimators, E1, E2, D1, D2 on the interval $[0, \tau]$ are given for setting A (experimental arm) and B (control arm) (see legend) as a function of τ. The bias is computed as the estimated RMDoR based on the four estimators minus the one based on daily scans (no interval censoring). The grey vertical lines indicate the time-points the patients had a scan. In the first plot the red and green lines are not visible as they overlap with the blue and purple lines, respectively. In this simulation 100 patients per setting is used.

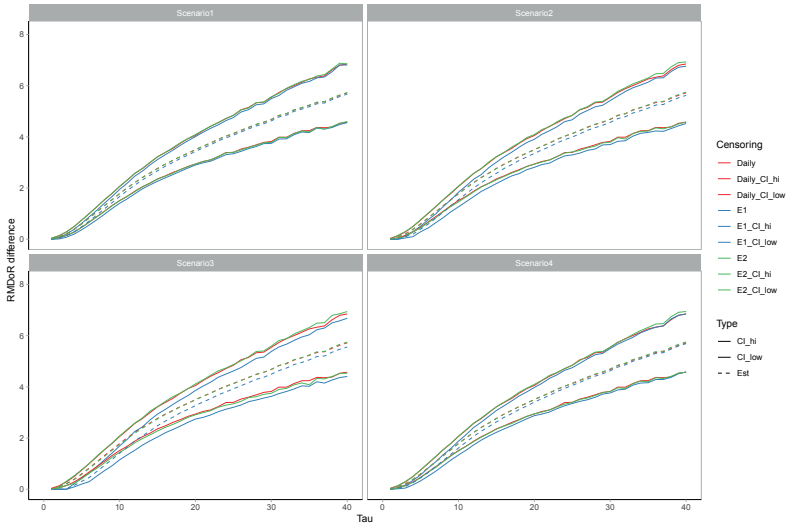


Figure 3.9: For the scanning schedules S1, ..., S4 and every estimator, the estimated difference of the RMDoR between the two arms (experimental arm minus control arm) are shown as a function of τ. Some curves are difficult to see, due to overlapping curves. In this simulation 100 patients per arm is used.

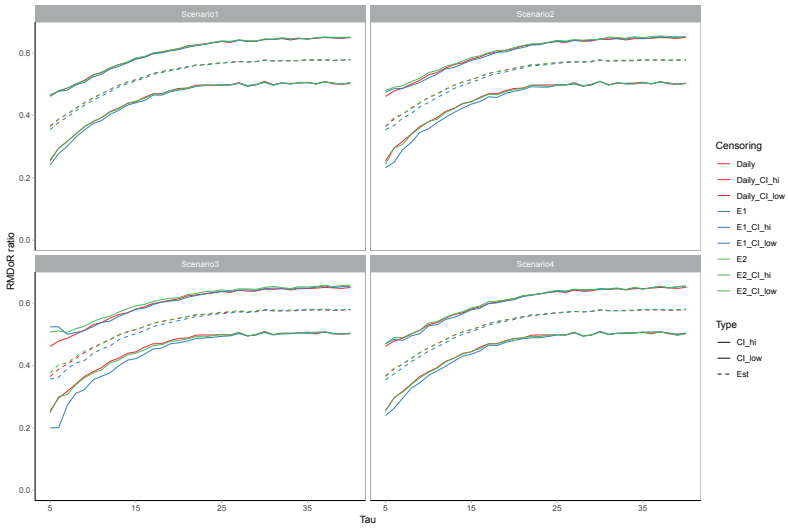


Figure 3.10: For the scanning schedules S1, ..., S4 and every estimator, the estimated ratio of the RMDoR (experimental arm/control arm) are shown as a function of τ. Some curves are difficult to see, due to overlapping of the curves. In this simulation 100 patients per arm is used.

4

Response rate estimation in single-stage basket trials: A comparison of estimators that allow for borrowing across cohorts

Authors: Antonios Daletzakis, Rutger van den Bor, Vincent van der Noort, Kit CB Roes

Original title: Response rate estimation in single-stage basket trials: A comparison of estimators that allow for borrowing across cohorts

Under review

Abstract

Therapeutic advancements in oncology have transitioned towards targeted therapy based on specific genomic aberrations. This shift necessitates innovative statistical approaches in clinical trials, notably in the emerging paradigm of master protocol studies. Basket trials, a type of master protocol, evaluate a single treatment across cohorts sharing a common genomic aberration but differing in tumor histology. While offering operational advantages, the analysis of basket trials introduces challenges with respect to statistical inference. Basket trials can be used to decide for which tumor histology the target treatment is promising enough to move to confirmatory clinical evaluation and can employ a Bayesian design to support this decision making. In addition to decision making, estimation of the cohort-specific response rates is highly relevant to inform design of subsequent trials. This study evaluates seven Bayesian estimation methods for basket trials with a binary outcome, contrasted with the (frequentist) sample proportion estimate, through a simulation study. The objective is to estimate cohort-specific response rates, with a focus on average bias, average mean squared error, and the degree of information borrowing. A variety of scenarios are explored, covering homogeneous, heterogeneous and clustered response rates across cohorts. The performance of the evaluated methods shows considerable trade-offs in bias and precision, emphasizing the importance of method selection based on trial characteristics. Berry's method excels in scenarios with limited heterogeneity. No clear winner emerges in a more general scenario, with method performance influenced by the amount of shrinkage towards the overall mean, bias and the choice of priors and tuning parameters in more complex settings. Challenges include the computational complexity of methods, the need for careful tuning of parameters and prior distribution specification, and the absence of clear guidance on their selection. Researchers should consider these factors in designing and analyzing basket trials.

4.1 Introduction

Therapeutic approaches in oncology have shifted from conventional chemotherapy to targeted therapy on a specific genomic aberration across different cancer types. The progress in the field of precision medicine required a different statistical approach to clinical trials. Master protocol studies appear to become increasingly common in practice, particularly in non-randomized exploratory phase I/II research. [12] [11] An example of a master protocol study is the basket trial design, the focus of this paper. In such trials, the same treatment is evaluated in multiple cohorts of patients with different tumour histologies that share the same genomic aberration.

Basket trials can be seen as a series of single-arm trials (usually designed as either single or two-stage) performed in distinct histology-based patient cohorts. An operational advantage of this design compared to conducting multiple individual trials for each tumor histology, is that it requires only a single protocol, a single database, and a single medical ethical review board application. As such it results in time and cost savings. With respect to the analysis of studies with a basket trial design, a straightforward inference strategy is to analyse each cohort separately, or to pool the data and provide a total estimate for the patients regardless of the patient's tumour histology. Either pooling or independently analyzing trial results in practical applications may not necessarily represent the optimal choice in all cases. When there is high heterogeneity the pooling estimate introduces type I error inflation and in low heterogeneity the independent analysis lacks of statistical power compared to alternative methods. A third option is to use a procedure which, like independent analyses, performs inference on individual cohorts, but which allows for borrowing information across cohorts. Such methods may provide advantages in terms of statistical efficiency. (see, e.g. Pohl et al. [14] for a comprehensive review).

Methods or designs that allow for borrowing information are typically based on Bayesian procedures and assume that the primary outcome of interest is the response rate: the fraction of patients in each cohort who showed, upon treatment, a clinically significant shrinkage in their tumor volume, a common endpoint in early phase oncology trials. Although the primary focus of the methods review in current article is on decision-making, posterior point estimates of the cohort-specific response rates can be derived as well. Such estimates are of relevance to plan subsequent trials or to provide estimates to support benefit-risk assessment and communicate expected effects to patients. While operating characteristics of the decision-making process are important as well, the focus of the current simulation study is on the performance (bias, MSE) of the response rate estimator derived from the posterior distribution used in the Bayesian procedures.

An overview of the performance of estimators based on seven Bayesian analysis methods that allow for borrowing is presented for a range of scenarios. Additionally, we investigate the influence of prior distribution choice on the performance of Bayesian estimators. All the estimators are applied in a variety of scenarios all considering parallel single stage cohorts. A range of scenarios is addressed, encompassing homogeneous, heterogeneous, and varied levels of response rate distribution across cohorts.

In section 4.2.1 the methods used in this paper are presented in detail. In section 4.2.2 the setting, the simulation methodology and the explored scenarios are discussed, including the approach to compare the results. In section 4.3, the results from our simulation study are presented as the evaluation of the methods in terms of bias, MSE and amount of information borrowing. The methodology and the limitations of the comparative evaluation are discussed in the Discussion section 4.4 of this paper, where we also provide suggestions for trialists.

4.2 Methods

4.2.1 Estimators

The objective of estimation in the trial is to produce, for each cohort i, an estimate $\hat{p}_i$ of the *true response rate* p_i of that cohort, i.e. the probability of responding to the treatment for a randomly chosen patient in the i^{th} cohort. Some methods assess p_i through the the log odds parameter θ_i defined as $\theta_i = \log(\frac{p_i}{1-p_i})$, or, equivalently, $p_i = \frac{1}{1+\exp(-\theta_i)}$.

In this paper, besides the cohort-specific sample proportion (the observed number of responders divided by the total number of subjects), seven Bayesian procedures to estimate cohort-specific response rates allowing for borrowing are evaluated:

Estimator based on Berry et al.[54]

Berry et al. [54] discussed the use of a Bayesian hierarchical model (BHM) as a method that allows for borrowing information across all cohorts. The BHM assumes that the log odds parameters θ_i are exchangeable between cohorts,* meaning that all cohorts follow the same distribution i.e. $N(\mu, \sigma^2)$. The hyper-priors, for μ and σ^2, are defined as a normal distribution $N(\mu_0, \sigma_0^2)$ and an inverse gamma distribution $\sigma^2 \sim IG(\lambda_1, \lambda_2)$ respectively.

Estimator based on Neuenschwander et al.[55]

The EXNEX [55](exchangable-nonexchangable) method is a BHM method that extends the conventional Berry's BHM by relaxing the assumption of all cohorts being exchangeable. Less information is being borrowed between non-similar cohorts. The log odds parameter θ_i for each cohort follows either a distribution which allows to exchange information, EX: $\theta_i \sim N(\mu_0, \sigma_0^2)$ with probability w, or a distribution that is non-exchangeable with a probability 1- w, NEX: $\theta_i \sim N(m_i, v_i^2)$. The hyper-parameters employed in the NEX component, namely m_i, v_i^2 and w, are fixed. In this study, priors and parameters were specified in accordance with the recommendations of the EXNEX [55] authors. The hyper-parameters used in the EX component are $\mu_0 \sim N(0, 10)$ and $\sigma_0^2 \sim$ half-normal(1). For the purposes of this paper the 'bhmbasket' R package is used for the calculation of the estimate.

Estimator based on Psioda et al.[57]

Here, response rates are estimated following the procedure in Psioda et al. [57], who propose a Bayesian model averaging approach. All possible models (ranging from the most parsimonious model in which all estimates are constrained to be equal to the most complex model in which all estimates are allowed to differ) are assigned a prior probability of being true. In addition, a beta prior is used for the response rate estimate in each cohort. Based on the observed data, the posterior model probabilities and model-specific posterior distributions for the response rates are determined. Cohort-specific response rates are calculated as the weighted average of the mean of the model-specific beta posterior distributions, with weights equal to the posterior model probabilities. Estimates were obtained using the function 'bma' (version 0.1.2) in the R package 'bmabasket'. We use the default parameter specifications, with the exception of the prior for the response rates, which are defined to be uniform (Beta(1,1)), instead of weakly informative.

*The methods Berry et al. [54], EXNEX [55](exchangeable-non exchangeable) and Jin et al. [56] model the response rate p_i using the log odds transformation. Instead of modelling p_i directly, they model the distribution of $\theta_i = logit(p_i) = log(\frac{p_i}{1-p_i})$.

Estimator based on Fujikawa et al. [58]

The Fujikawa et al. [58] method is a Bayesian approach that borrows information across cohorts based on the similarity of their response rates. First, a uniform ($Beta(1,1)$) prior is used for the response rate in each cohort, which is updated based on the observed data. The pairwise similarity between the resulting posterior Beta distributions is then determined using the Jensen-Shannon divergence. If the similarity exceeds a given pre-specified threshold (denoted by τ), the posteriors are 'combined' (i.e. borrowing will take place, by updating the parameters of the posterior distribution for a given cohort as a 'weighted average' of the initial posterior with all the other cohorts with similar effect size). We use the means of the possibly updated posterior distributions as the point estimates for the response rates.

Estimator based on Jin et al. [56]

The Bayesian hierarchical model with a correlated prior (CBHM) proposed by Jin et al. [56] is a method that allows borrowing more information between possibly homogeneous cohorts and less when the treatment effect seems heterogeneous. For the log odds parameter θ_i, it is assumed that it is specified as: $\theta_i = \theta_0 + \eta_i + \epsilon_i$. The η_i are cohort-specific effects which follow a multivariate normal distribution with correlation matrix Ω, $\eta_i \sim MVN(0, \sigma^2\Omega)$. The similarity between two cohorts is identified in the Ω matrix, a correlation function is generated by the pairwise distance measures d_{ij}. Three different distance measures are considered in the original paper (the Kullback-Leibler distance (KL), the Hellinger (H) distance and the Bhattacharyya (B) distance). For the purposes of this paper, we will use the H distance.

Estimator based on Chen & Lee [59]

Chen & Lee[59] proposed a Bayesian cluster hierarchical model. A two-step procedure, first the Chinese restaurant process (CRP) [60] identifies the partitioning into clusters of each cohort using a non-parametric Dirichlet process mixture model (DPM). The values of the clustering matrix C_{ij} are the proportion of two cohorts being classified in the same cluster. The second step is to use a Bayesian hierarchical model to estimate the log odds of the response rate $\theta_i \sim N(\mu_1, \frac{1}{\tau_1 m_i})$, given the cluster structure. The m_i in this model is set to C_{ij} for the specific subgroup i. A hyper-prior for the mean μ_1 follows a normal distribution $N(\mu_2, \frac{1}{\tau_2})$, a hyper-prior for the precision controlling the amount of borrowing between cohorts i and j, $\tau_1 \sim Gamma(\alpha_1, \beta_1)$ and μ_i is the respective model indicated by the C_{ij} matrix. The variance of the hyper-prior was explored by Chen & Lee [61], $\tau_2 = 0.1$. The clustering matrix probability of subgroup i and j to share information influences the variance of the θ_i distribution. As the similarity value increases, the information sharing becomes larger and the posterior distribution variance decreases. The 'BCHM' R package is used for the response rate calculation given the default parameters proposed by the authors. (More details in Appendix B)

Estimator based on Liu et al.[62]

Liu et al. [62] propose a method that evaluates the probabilities of all possible models (referred to as 'partitions'), assigning a prior probability to each partition differently than Psioda et al. [57]. The choice of prior is carefully discussed, with a parameter, delta, introduced to determine the level of influence each model exerts. The authors suggest values of 0 (uniform prior), 1, and 2 for delta. Unlike the Psioda method, which uses a weighted average, Liu et al.[62] select the most probable partition to compute the pairwise similarity matrix among cohorts. This matrix is then used to calculate the parameters of the Beta posterior distribution. In the local multiple exchangeability

model (local MEM), this similarity matrix is used to determine which baskets are grouped together based on the highest posterior probability partition. Information borrowing is then carried out locally within these identified groups. Specifically, baskets that belong to the same block, as defined by the selected partition, are treated as exchangeable, and data from these baskets are used to update the parameters of their Beta posterior distributions. This localized borrowing approach allows for precise information sharing while respecting differences between dissimilar baskets.

4.2.2 Simulation study

Setup

In the simulation study, we construct a basket trial with six different cohorts, which we consider to be a realistic choice. For simplicity, all cohorts of the study are assumed to be single-stage and have equal size. Limited to no prior knowledge is assumed to be available. The total number of patients per cohort is denoted by n_i and the number of responses is indicated by r_i.

Scenarios

The scenarios in which we evaluate the response rate estimators differ in the distribution of the true response rates p_i across cohorts and the number of patients per cohort. We classify these scenarios into two types: homogeneous and heterogeneous (table 4.1). The choice of the different scenarios is based on the methodological aspects of the methods, but also extreme practical examples, like the KEYTRUDA [63] trial, lead us to consider a broad range of scenarios.

- **Homogeneous Scenarios:**

  In homogeneous scenarios, all cohorts have similar or identical true response rates. These scenarios are designed to evaluate how well the estimators perform when there is little to no variation between cohorts. Specifically:

  Scenarios 1.A.1 to 1.A.3: The true response rate is the same across all six cohorts, with values set at 0.1, 0.3, or 0.5, respectively.

- **Heterogeneous Scenarios:**

  In heterogeneous scenarios, the cohorts have more distinct response rates, representing a wider range of treatment effects across the different groups. These scenarios are designed to test the ability of the estimators to handle substantial variability:

  Scenarios 1.B.1 to 1.B.4: Small variations in true response rates are introduced between cohorts. In scenarios 1.B.1 and 1.B.2, the response rates differ slightly (by 0.025 and 0.05 respectively) around an overall mean of 0.5. In scenarios 1.B.3 and 1.B.4, the overall mean is 0.3, with variations of 0.025 and 0.05 respectively.

  Scenarios 2.A.1 to 2.D.3: These scenarios represent situations where two distinct groups of cohorts have different response rates. Scenarios 2.A.1 to 2.B.3 response rate with difference of 0.2. Scenarios 2.C.1 to 2.C.3 represent a larger response rate difference of 0.4, while scenarios 2.D.1 to 2.D.3 consider an even greater difference of 0.6 between the two groups.

Scenarios 3.A.1, 3.A.3, 3.B.1, and 3.B.3: In scenarios, 3.A.1, 3.A.3, 3.B.1, 3.B.3, we assume three different response rates, where four of the cohorts have the same response rate, while the remaining two cohorts have different rates. Scenarios 3.A.2 and 3.B.2 describe cases where three pairs of cohorts each have the same response rate, resulting in three different groups.

The responses of subjects in each cohort were generated from a binomial distribution with true probability p_i, where $i \in \{1, 2, ..., 6\}$. The number of subjects per cohort was assumed to be $N = \{10, 20, 30, 100\}$.

Details concerning the specification of prior distribution parameters and tuning parameters of the seven estimators are provided in Appendix B. R-code is available as online supplementary material.

Scenarios	Coh A	Coh B	Coh C	Coh D	Coh E	Coh F
1.A.1	0.1	0.1	0.1	0.1	0.1	0.1
1.A.2	0.3	0.3	0.3	0.3	0.3	0.3
1.A.3	0.5	0.5	0.5	0.5	0.5	0.5
1.B.1	0.4375	0.4625	0.4875	0.5125	0.5375	0.5625
1.B.2	0.375	0.425	0.475	0.525	0.575	0.625
1.B.3	0.2375	0.2625	0.2875	0.3125	0.3375	0.3625
1.B.4	0.175	0.225	0.275	0.325	0.375	0.425
2.A.1	0.3	0.5	0.5	0.5	0.5	0.5
2.A.2	0.3	0.3	0.3	0.5	0.5	0.5
2.A.3	0.3	0.3	0.3	0.3	0.3	0.5
2.B.1	0.1	0.3	0.3	0.3	0.3	0.3
2.B.2	0.1	0.1	0.1	0.3	0.3	0.3
2.B.3	0.1	0.1	0.1	0.1	0.1	0.3
2.C.1	0.1	0.5	0.5	0.5	0.5	0.5
2.C.2	0.1	0.1	0.1	0.5	0.5	0.5
2.C.3	0.1	0.1	0.1	0.1	0.1	0.5
2.D.1	0.1	0.7	0.7	0.7	0.7	0.7
2.D.2	0.1	0.1	0.1	0.7	0.7	0.7
2.D.3	0.1	0.1	0.1	0.1	0.1	0.7
3.A.1	0.1	0.4	0.7	0.7	0.7	0.7
3.A.2	0.1	0.1	0.4	0.4	0.7	0.7
3.A.3	0.1	0.1	0.1	0.1	0.4	0.7
3.B.1	0.1	0.4	0.9	0.9	0.9	0.9
3.B.2	0.1	0.1	0.4	0.4	0.9	0.9
3.B.3	0.1	0.1	0.1	0.1	0.4	0.9

Table 4.1: Scenarios used in the explored simulations. The first 3 scenarios are the Homogeneous scenarios and the rest are referred to the Heterogeneous scenarios.

Evaluation criteria

The estimators will be compared based on the average absolute bias and average MSE, as defined in Table 4.2.

An additional measure used to provide further insight in the results, is the Shrinkage to the total mean, which is defined as the difference between the maximum and the minimum estimated response rate, divided by the simulated max and min. We use the form as presented in table 4.2, such that the closer this value is to 1, the greater the shrinkage. When the shrinkage is close to 0, it indicates that the methods are not borrowing much information.

		Formula
MeanAbsBias	$=$	$\frac{\sum_{i=1}^{6} \lvert E(\hat{p}_i) - p_i \rvert}{6}$
$MeanMSE$	$=$	$\frac{\sum_{i=1}^{6} [Bias(\hat{p}_i)^2 + Var(\hat{p}_i)]}{6}$
$Shrinkage$	$=$	$1 - \frac{max(E(\hat{p_1}),\ldots,E(\hat{p_6})) - min(E(\hat{p_1}),\ldots,E(\hat{p_6}))}{max(p_1,\ldots,p_6) - min(p_1,\ldots,p_6)}$

Table 4.2: Evaluation criteria

4.3 Results

4.3.1 Homogeneous scenario

In the homogeneous scenarios with exactly equal true response rates across the baskets (scenarios 1.A.1 to 1.A.3, fig 4.1 all Bayesian estimators except Chen & Lee show on average positive bias for response rates 0.1 and 0.3, but not when the response rate is 0.5. The Chen & Lee shows negative bias when the true estimates are lower than 0.5. Among the Bayesian estimators, Berry's BHM [54] shows the smallest bias and average MSE, regardless of the sample size per basket. The EXNEX and Jin methods are slightly shifting towards 0.5. When the sample size is small, the average absolute bias for some of the other estimators, particularly those based on Liu and Fujikawa, appears quite substantial towards 0.5 when the true common response rate is 0.1, but overall bias decreases to negligible when the true response rate gets closer to 0.5.

To determine whether this effect is actually due to the prior choices, we conducted additional simulations, (Appendix C, fig 4.3,4.4) in which the parameters of each method were tuned to have a prior mean of 0.3. The results indicate that the prior mean plays a role and influences whether the estimates overestimate or underestimate the effect, towards to the chosen prior mean. Berry's method is not affected by the prior, as well as Chen & Lee. The EXNEX and Jin methods are affected slightly. The Psioda method is affected, but less than the Fujikawa and Liu methods.

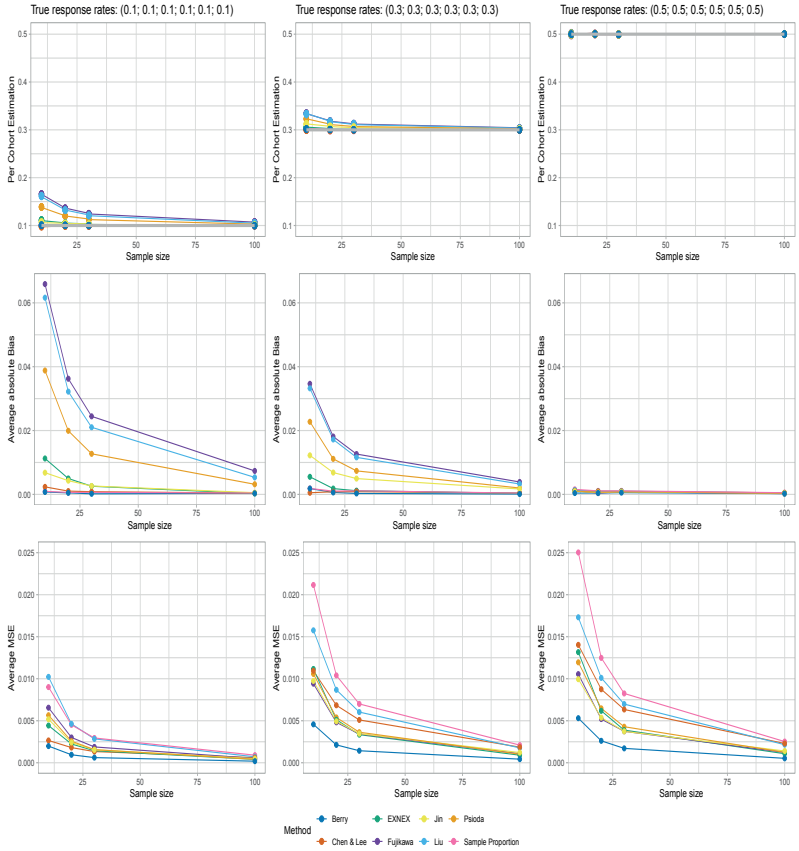


Figure 4.1: *Estimates on scenarios of the same true response rate 1.A.1 to 1.A.3 in 4 sample size points, 10, 20, 30, 100 (patients per cohort). In the 1st row the estimates are presented, in the second and third, the average absolute bias and the average MSE of the respective scenarios. The gray lines in the first row reflect the true response rates. The prior distribution in this graph is having a mean of the 0.5*

4.3.2 Heterogeneous Scenario

If we allow for some heterogeneity in the true RRs of the six cohorts (scenarios 1.B.1 to 1.B.4, fig 4.8 and 4.9), the observed bias of the different Bayesian estimators increases. The average mean squared errors of the different estimators overall increase and become more similar.

In fig 4.8, in the first row the estimates can again be seen to be biased towards 0.5, regardless of the variation level. In fig 4.6 it can be seen that different true RRs across the 6 cohorts centered around 0.5, that the estimates shift to this mean of 0.5 of the cluster. In fig 4.9, the mean of true RRs of the cohorts is 0.3, and most of the estimates show shift towards 0.5, which was not observed in scenarios where the overall cluster mean of the true RRs is 0.5.

In these scenarios, Berry's BHM estimator has the lowest MSE when the response rate is less heterogeneous, compared to the other methods. However, with increasing variability, the average absolute bias appears relatively large, even with a large sample size. Also for other estimators, bias still appears to be present even with samples sizes as large as 100 per cohort.

When considering scenarios involving the grouping of cohorts (i.e., two or three distinct groupings), all methods effectively capture the structure (fig 4.2). As anticipated, the sample proportion has the smallest average absolute bias (theoretically 0), but its average Mean Squared Error (MSE) is among the highest (see fig 4.2). In this setting, Berry's Bayesian Hierarchical Model (BHM) estimator demonstrates the largest average absolute bias across most scenarios. Estimators following Fujikawa and Liu's approaches tend to shrink more towards the prior of 0.5, compared to other estimates. Jin's estimator shrinkage is small. EXNEX, Psioda, and Chen & Lee estimates also show a tendency towards the mean of 0.5, but with a notable exception for the Chen & Lee estimator (see fig 4.6): it uniquely shifts towards 0 when the estimated total mean is below 0.5 and towards 1 when above 0.5.

In scenarios as in fig 4.11, where the actual response rate of five out of 6 cohorts is equal to 0.5 and one cohort deviates, the methods that tend to shift estimates towards 0.5 (such as Fujikawa's) perform better in terms of MSE. Berry's BHM estimator appears favourable with respect to average MSE when the heterogeneity is small (see fig 4.10, 4.9). However, as the heterogeneity increases (see fig 4.11, 4.12, 4.7), Berry's estimator's MSE grows relative to others. The observed patterns remain similar across all scenarios.

The diversity across different cohorts impacts the degree of shrinkage to the overall mean observed in our analysis. Specifically, as the heterogeneity between cohorts increases, we notice a corresponding decrease in shrinkage (see fig 4.5). Berry's method consistently shrinks more towards the overall mean, irrespective of the level of heterogeneity. In contrast, Liu's method exhibits the least shrinkage across all scenarios. The Chen and Lee method tends to shrink towards the overall mean, similar to most methods when variability is low, but more prominently than others in scenarios with large between cohort heterogeneity. Overall, shrinkage decreases with greater cohort heterogeneity. Additionally, we observe that shrinkage tends to diminish as the sample size increases, although certain methods—like Berry's—maintain high shrinkage under conditions of small heterogeneity.

As described in the homogeneous scenarios, the choice of prior has an impact on the results, particularly by shifting estimates towards the prior mean. This effect is evident in fig 4.4, which contrasts the results when using a prior mean of 0.3, compared to the prior mean of 0.5 shown in fig 4.2. Most estimates shift accordingly to the homogeneous scenarios, illustrating the sensitivity of each method to the choice of prior.

In general, the differences become more apparent with smaller cohort sample sizes, though no consistent ranking emerges among estimators. As sample sizes increase, estimators generally become more accurate and similar. Specifically, most methods converge to the true RR with cohort sizes of 100 or more. The advantages of these methods over the sample proportion in terms of lower MSE diminish in larger sample contexts.

A concise summary of these results can be found in Table 3.

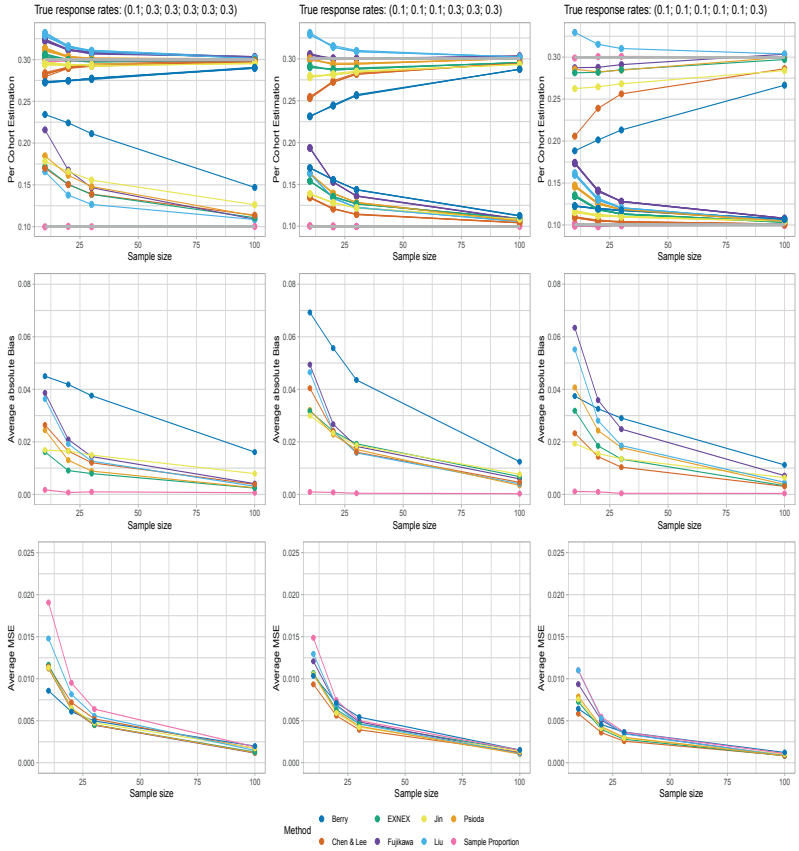


Figure 4.2: *Estimates on scenarios 2.B.1 to 2.B.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE. The gray lines in the first row reflect the true response rates. The prior distribution in this graph is having a mean of the 0.5*

4.4 Discussion

In this paper, we present a comprehensive review of Bayesian methodology for a basket-type design with single arm cohorts from the perspective of estimation. We evaluated various Bayesian methods and included the cohort sample proportion for reference. We limited the evaluation to the setting of parallel single-stage cohorts using identical sample sizes. The main objective of this paper is to assess how these methods perform with respect to estimation of success proportions for the different cohorts in the basket trial. To ensure fair comparisons, we use flat or weakly informative prior distributions, following the original author's recommendations. The evaluation of the estimates considers frequentist properties including mean absolute bias and average mean squared error (MSE) and shrinkage metrics. The motivation for this research is the limited availability of research results that evaluate estimation properties when information borrowing is applied, whereas such estimation is particularly relevant for early phase basket trials that inform subsequent larger (confirmatory) trials.

In a basket trial, statistical borrowing from other cohorts to improve estimation of the response rate for an individual cohort should essentially serve to improve the estimation compared to the cohort (sample) proportion. It is typically considered when the cohorts target rare conditions, and thus sample sizes per cohort are relatively small. The rationale of the basket trial (e.g. shared molecular target) typically provides justification for such borrowing. Even when such (mechanistic) justification is strong, heterogeneity in response rates between cohorts usually cannot be excluded a priori. Therefore, key desirable properties of Bayesian methods for estimation in this setting are twofold: improved precision at cohort level when heterogeneity is limited or absent, and sufficiently sensitive to adapt (i.e., limit bias at cohort level) when heterogeneity between cohorts is clearly present.

When put against these criteria, we conclude that there is no clear winner among the Bayesian methods in terms of optimal average MSE and average absolute bias across the scenarios evaluated. (See also table 4.3) Berry's et al. [54] demonstrates the smallest average MSE and average absolute bias when the true response rate belongs to the homogeneous scenario. When a higher heterogeneity level is introduced there is no optimal choice. Except for large cohort sample sizes, overall bias and mean squared errors are relatively substantial in some cases, especially as heterogeneity increases. The methods differ in the amount of shrinkage and in the amount that the estimate is influenced by the prior. All the methods shrink towards the overall mean across the cohorts. Berry's et al. [54], Jin's et al. [56], and Chen & Lee [59] estimates shrink more towards the overall mean compared to other Bayesian estimates. Fujikawa et al. [58] and Liu et al. [62] seem to be pulled towards 0.5 compared to the other estimates, as clearly seen in fig 4.6. Chen & Lee [59] estimate, on the other hand, differs from the other estimators as it appears to shrink towards 0 if the overall true response rate mean is smaller than 0.5 and towards 1 if this is larger is over 0.5. Given the set criteria Berry's et al. [54], Fujikawa et al. [58] and Liu et al. [62] methods are less preferred, and EXNEX, Psioda's, Jin's et al. [56], and Chen & Lee [59] methods are more suitable, with the EXNEX method to be more consistent than the other methods. Notably, the performance of the sample proportion is equal or superior compared to the Bayesian methods in terms of MSE when the heterogeneity increases, and of course it is unbiased. In this paper, the focus was on evaluating the methods under a setting where there is no formal prior information concerning the homogeneity and the heterogeneity structure that could be included in the estimation. Using more informative priors does fit within some of methods, which could lead to more precise outcomes.

The complexity of the methods introduces a potential limitation to the present work and applications. The tuning parameters needed for the different methods imply that there is substantial flexibility (hence heterogeneity) in their implementation, without clear a priory guidance to optimally set these parameters. We used default values and non-informative priors in a simple setting, but in practice other choices may be made. The non-informative prior distributions has a prior mean of 0.5, when we made a different prior mean choice, we see a different be-

haviour in the estimated results.

In our effort to (computationally) replicate these methods for our simulation study, we faced many difficulties in selecting the appropriate parameters and ensuring that our implementation was indeed exactly equal to the published results and scripts and also occasionally found discrepancies between the publicly shared script and the scientific paper (that were subsequently resolved). Researchers may face similar struggles in determining the method most suitable for each situation and ensuring an appropriate computational implementation. Additionally, fully understanding the methods and their implementation is not straightforward, with usually guidance lacking in addressing estimation.

Several simplifying assumptions were used in the simulation study. The choice for 6 cohorts of equal sample size with a single-stage design only covers a limited number of potential scenarios. In practical settings, it is common to suggest at least one interim analysis (e.g., following Simon's two-stage designs). In the single-stage setting, the sample proportion is an unbiased estimator, hence it was included as reference in our study. The proposed Bayesian methods do allow for interim analyses. Jin et al. [56] proposed the use of a single interim analysis, while Fujikawa et al. [58] and Psioda et al. [57] allow several interim futility assessments. Berry et al. [54] proposed an interim analysis when a certain number of patients are included (e.g., 10) and more assessments allowed after a pre-specified number of patients (e.g., 5). Simon et al. [64] considers the futility assessment after each observation. Neuenschwander et al. [55] and Chen & Lee [59] do not propose an interim analysis stage, but an interim analysis could be applied at any time point of the trial. We did not evaluate the resulting very broad range of possible scenarios, which may lead to some differences between the methods when interim analyses are implemented. However, as the number of cohorts with similar response rates investigated correspond to realistic practical settings and a range of sample sizes was explored, we do believe the present provides basis for an initial choice of methods.

The explored methods offer a potentially valuable tool for researchers in efficiently designing and analyzing basket trials. Estimators based on methods that allow borrowing of information across cohorts introduce bias, but are expected to have a smaller MSE, due to an increase in precision. However, there are many available options even within each method, with most methods requiring choices of tuning parameters in addition to priors for model parameters. A priory guidance on precise settings of these parameters for practical applications is challenging, which may limit the possibility to pre-specify the full estimation procedure. A simulation study, such as the one performed here, but targeted to a specific context of a particular study, could give a better insight.

Appendix A

Methods	Prior effect	Homogeneous scenario	Heterogeneous scenario	Shrinkage
Psioda	bias toward the prior mean	Small MSE, slightly biased estimate	Borrowing to the mean and bias to the prior	Moderate to small shrinkage in all scenarios
Berry	no effect	Smallest MSE, no bias	Bias estimate to the total mean	Extreme shrinkage to the total mean
EXNEX	limited - no influence	Small MSE	Shares less information when 1 vs 5 true RR basket, small MSE	Small shrinkage in small heterogeneity. Almost no shrinkage in high heterogeneity
Jin (CBHM)	Slight influence	Small MSE, slightly biased	Slight prior influence, Small MSE	Moderate shrinkage in small heterogeneity. Small shrinkage in high heterogeneity
Chen & Lee (BCHM)	No influence	Small MSE, almost unbiased, estimates to 0 if true RR $<$ 0.5 or towards 1 if true RR $>$ 0.5	Shrinkage to the total mean, the smallest MSE	Small shrinkage in small heterogeneity. Moderate shrinkage in high heterogeneity
Fujikawa	bias toward the prior mean	Bias to prior mean, small MSE	Extreme bias to the prior mean	Moderate shrinkage in small heterogeneity. Small shrinkage in high heterogeneity
Liu	bias toward the prior mean	Bias to prior mean, moderate MSE	Extreme bias to the prior mean	Almost no shrinkage regardless the heterogeneity level
Sample prop		The largest MSE	MSE close to the Bayesian methods, in highly heterogeneous trials smaller MSE	No shrinkage

Table 4.3: Results overview table

Appendix B

Prior and parameters choice

The choice of the prior is a hard task to handle, especially since the choice usually affects the estimation. We made a choice to use weakly informative or uninformative priors to compare the methods under the same rules, following the suggestions of the respective authors for uninformative priors to use in their method.

Jin et al. [56], Berry et al. [54], and EXNEX [55] do recommend somewhat informative priors involving pre-specified response rates q_0 and q_1 that correspond to inactive and active treatments respectively. Many of the methods are originally designed for a setting in which the goal of the study is to decide whether the treatment is active or not (represented by posterior probabilities of response rate $\geq q_1$ and $\leq q_0$ respectively). However, in our simulation study we focus on estimation of the response rates rather than decision making and use uninformative priors as specified below for each method.

- The ***Berry*** et al.[54] method, proposes a strategy to provide non informative choices of hyper-parameters. For this case, we set the hyper-parameters as follows:

$$\begin{aligned} \mu &\sim N(\mu_0, \sigma_0^2) = N(0, 100) \\ \sigma^2 &\sim IG(\lambda_1, \lambda_2) = IG(0.0005, 0.00005) \end{aligned}$$

- The ***EXNEX***[55] following the suggestions in the original paper we use a single exchangeability distribution of the EXNEX. The probabilities of each distribution is fixed, in the simple EXNEX model, the EX part is $p_i = 0.5$, and the NEX is $1 - p_i = 0.5$. A normal prior distribution is proposed and the mean can reflect the expectation of the researcher in the logit scale. In order to make a similar choice with the methods that uses Beta(1,1) prior and the mean of this prior is the 0.5, we set the mean of the EXNEX prior to be 0, which is the analogous of 0.5 in the logit scale.

$$\begin{aligned} \theta_i^{EX} &\sim N(\mu_0, \sigma_0^2) \\ \mu_0 &\sim N(0, 10) \\ \sigma_0^2 &\sim half - normal(1) \\ (p_i, 1 - p_i) &= (0.5, 0.5) \\ \theta_i^{NEX} &\sim N(m_i, v_i) \\ m_i &= 0 \\ v_i &= 10 \end{aligned}$$

- ***Psioda's***[57] method in the original paper is to set the a weakly informative Jeffreys prior $Beta(0.5, 0.5)$. For the purpose of this paper, we use the uninformative $Beta(1, 1)$. A prior model is set for the all possible models. The default settings of the package allow us to have a weakly informative prior where the model probabilities are giving greater weight to the more complicated models (less borrowing allowed).

- ***Fujikawa***-like [58] method estimates uses a beta prior $Beta(a_j, b_j) = Beta(1, 1)$. and the amount of borrowing is tuned by the parameter of $\tau = 0.5$.

- ***Jin*** et al.[56] made an extensive simulation study to address the effect of weakly informative prior distributions on the proposed design, specifying each set of priors for the respective distance measure of the correlation matrix.

  Hellinger (H) distance also introduces an exponential correlation, where the $\phi = Gamma(1.5, 1)$ has a different prior mean. Similarly the parameters are specified by:

$$\begin{aligned} \theta_0 &\sim N(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim IG(c_{\sigma^2}, d_{\sigma^2}) = IG(0.01, 0.01) \\ \tau^2 &\sim IG(c_{\tau^2}, d_{\tau^2}) = IG(0.01, 0.01) \\ \sigma_0^2 &\sim IG(c_{\sigma_0^2}, d_{\sigma_0^2}) = IG(0.1, 0.1) \end{aligned}$$

- ***Chen & Lee*** [59] present the choice of the prior and the parameters used in their paper in detail. For the classification model a choice of non-informative conjugate normal distributed prior $\mu = 0.2$ and $\sigma_0^2 = 10$ is used to calculate the posterior probability of the true response rate. The parameter $\alpha = 10^{-60}$ and $\sigma_d^2 = 0.001$ choice can affect the cluster number, which are used in the Dirichlet process (DP). The hyper-prior parameters calculated in order to propose a non-informative prior choice.

$$\begin{aligned} \theta_i &\sim N(\mu_1, \frac{1}{\tau_1 m_i}) \\ \mu_1 &\sim N(\mu_2, \frac{1}{0.1}) \\ \tau_1 &\sim Gamma(50, 10) \end{aligned}$$

- ***Liu*** [62] proposed an uninformative prior for the posterior distribution which follows a $Beta(1, 1)$. The level of leverage that each of the complicated models has, is chosen by the parameter δ. We choose to use the $\delta = 0$, so each model is weighted equally.

Appendix C

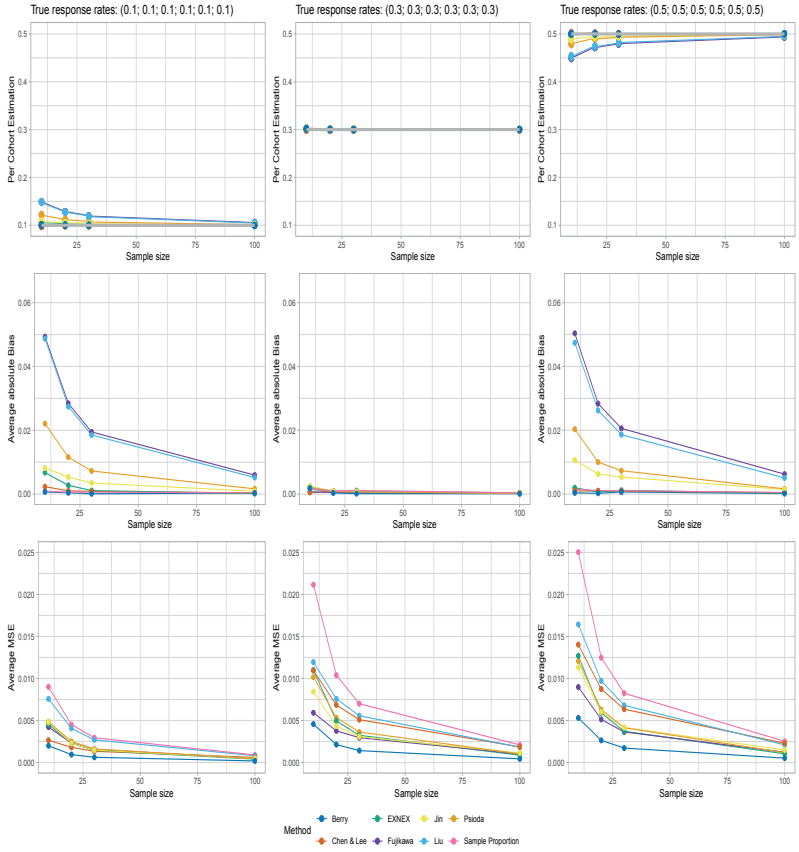


Figure 4.3: *Estimates on scenarios of the same true response rate 1.A.1 to 1.A.3 in 4 sample size points, 10, 20, 30, 100 (patients per cohort). In the 1st row the estimates are presented, in the second and third, the average absolute bias and the average MSE of the respective scenarios. The gray lines in the first row reflect the true response rates. The prior distribution in this graph is having a mean of the 0.3*

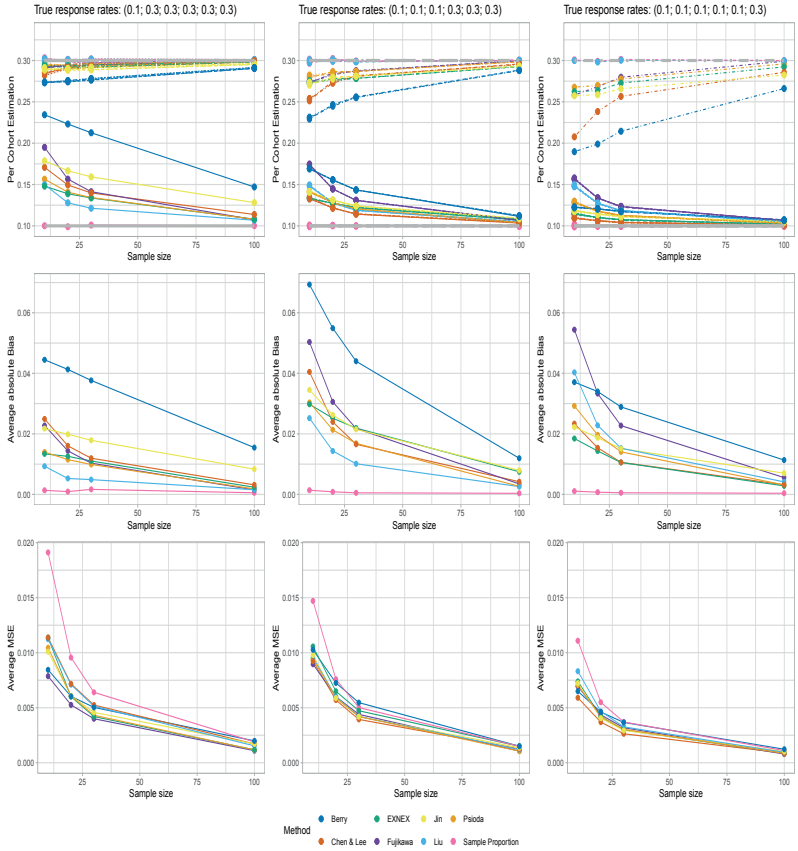


Figure 4.4: *Estimates on scenarios 2.B.1 to 2.B.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE. The gray lines in the first row reflect the true response rates. The prior distribution in this graph is having a mean of the 0.3*

Appendix D

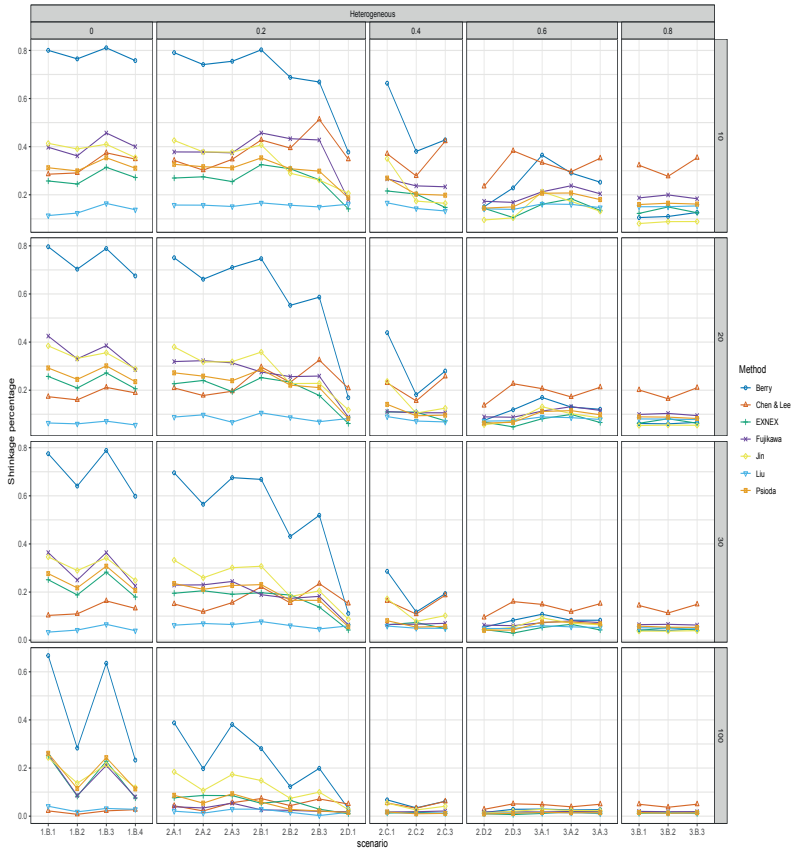


Figure 4.5: *Shrinkage of the estimates. Compare the range of the maximum and the minimum distance of the estimates with the true distance. Closer to 1 on y axis means the estimator shrinkage to the total mean is extreme. On the contrary if it's closer to 0, then the borrowing information is limited.*

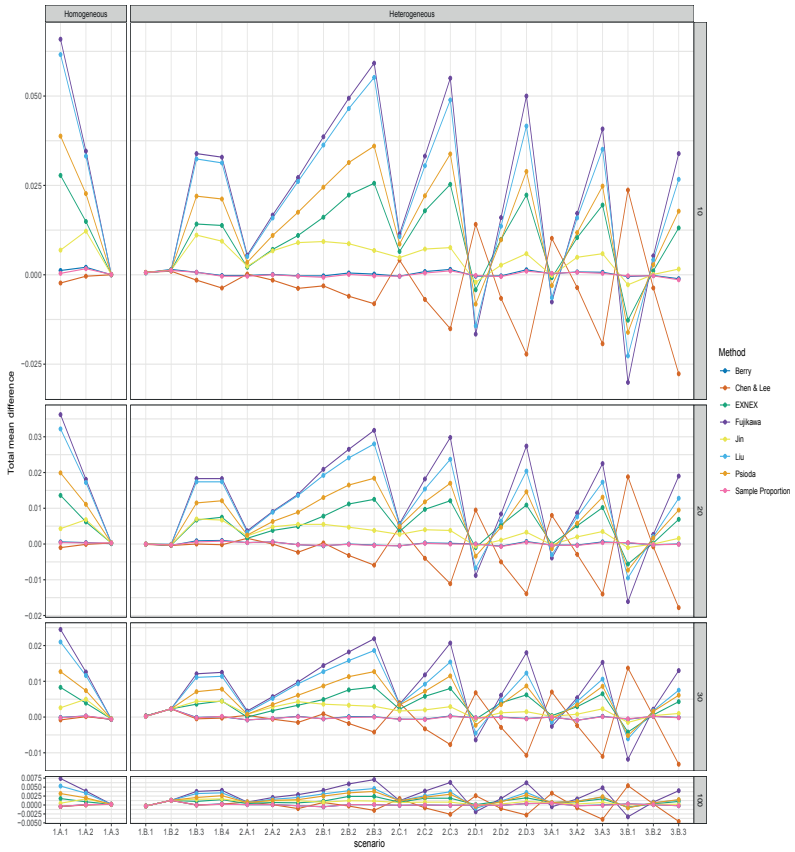


Figure 4.6: *Difference between the mean true RR of all cohorts with the mean of the simulated response rates in each method and every scenario. Each row indicates the different sample size points.*

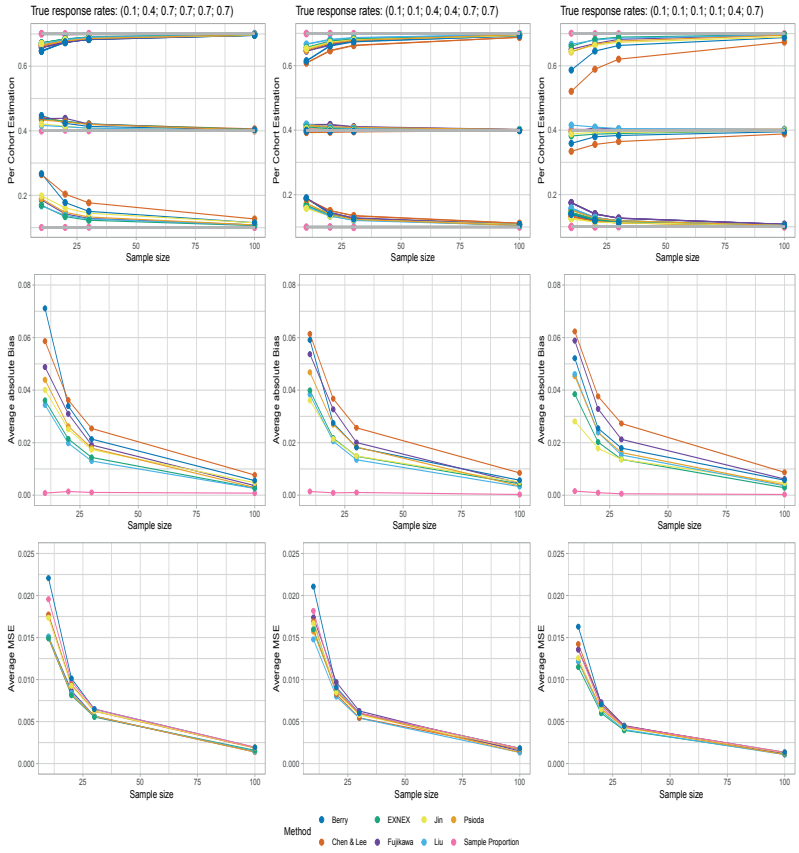


Figure 4.7: *Estimates on scenarios 3.A.1 to 3.A.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE. The gray lines in the first row reflect the true response rates.*

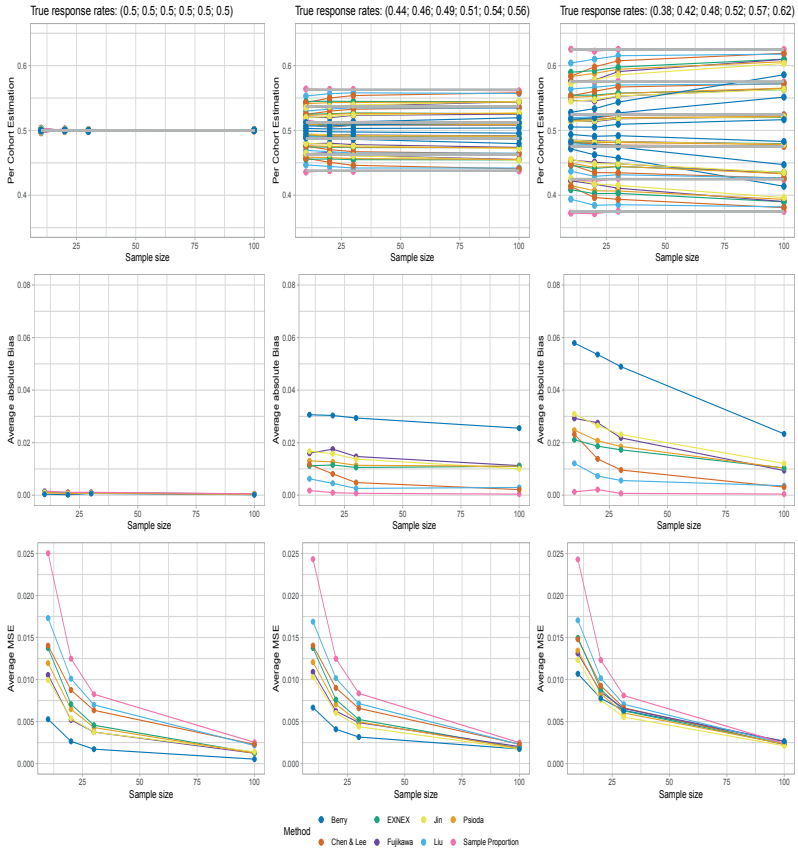


Figure 4.8: *Estimates on scenarios 1.A.3 to 1.B.2, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

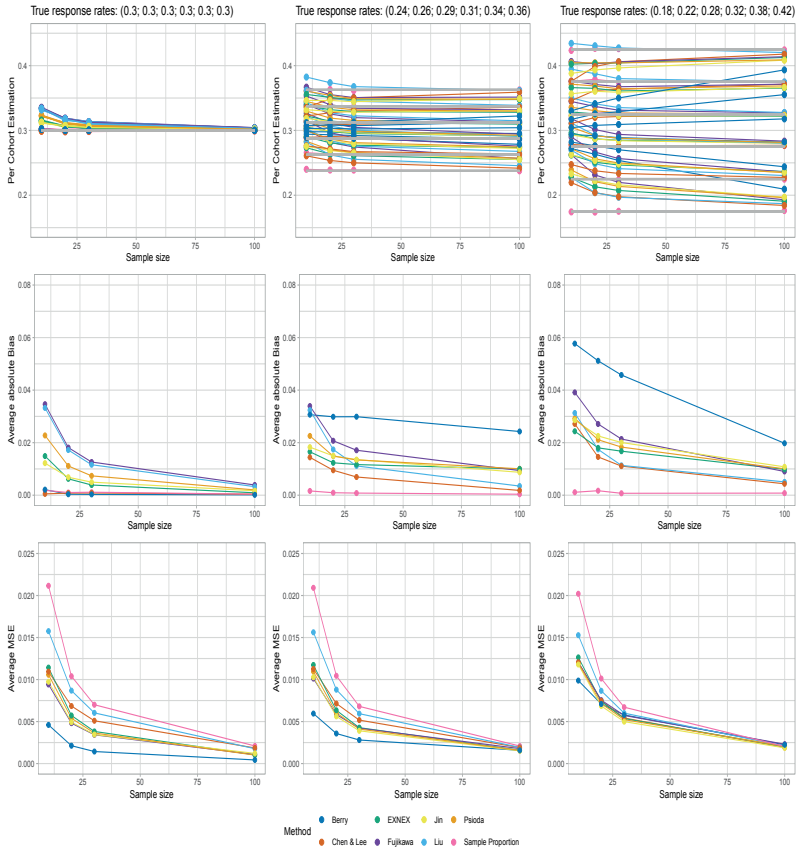


Figure 4.9: *Estimates on scenarios 1.A.2, 1.B.3 and 1.B.4, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

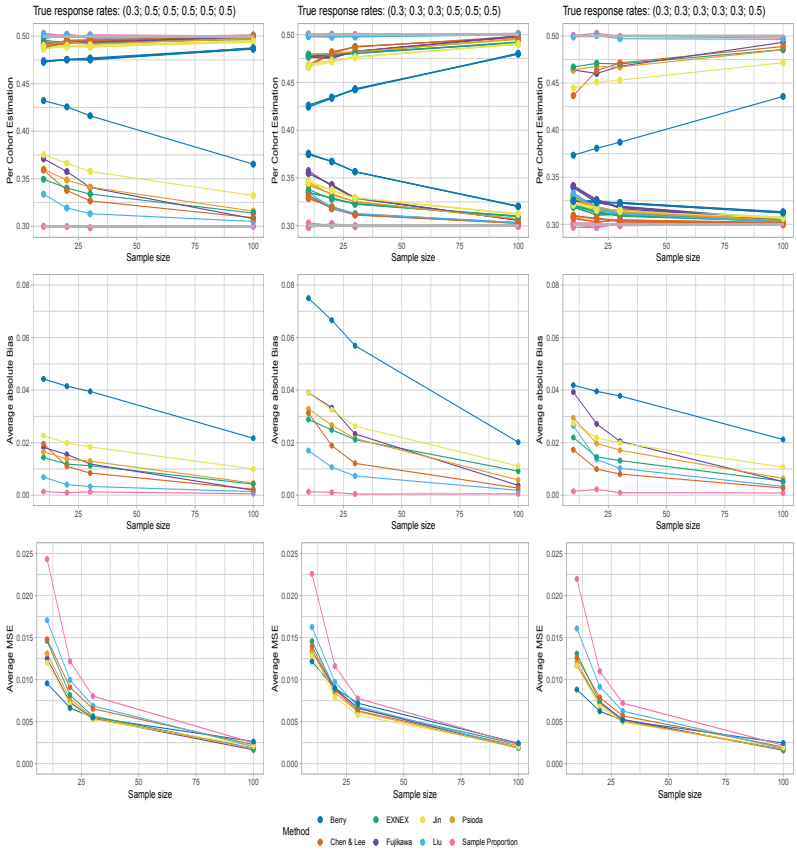


Figure 4.10: *Estimates on scenarios 2.A.1 to 2.A.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

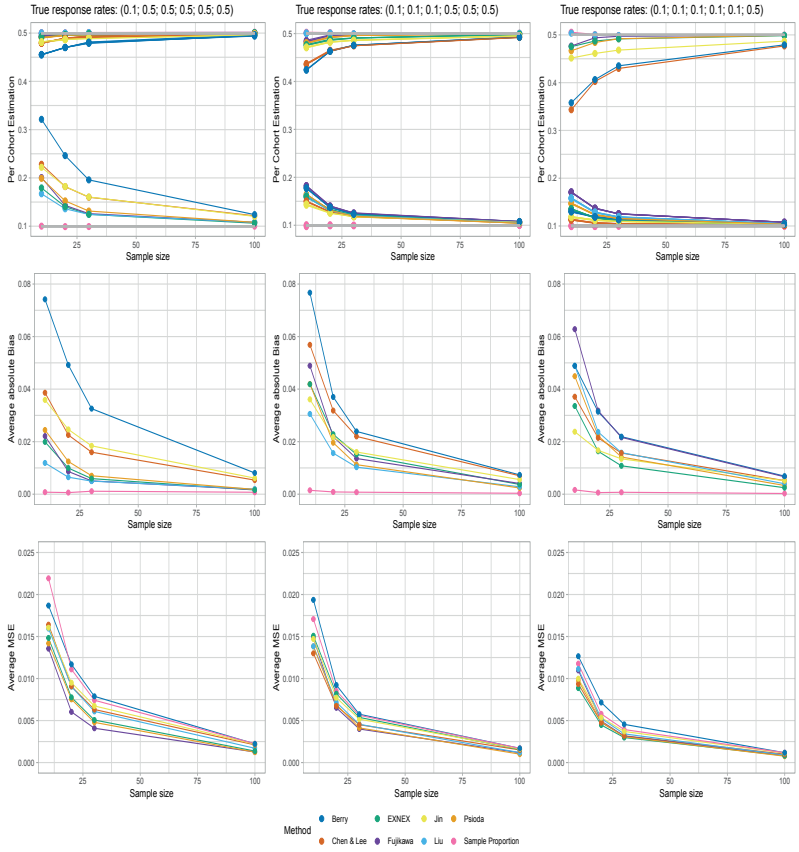


Figure 4.11: *Estimates on scenarios 2.C.1 to 2.C.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

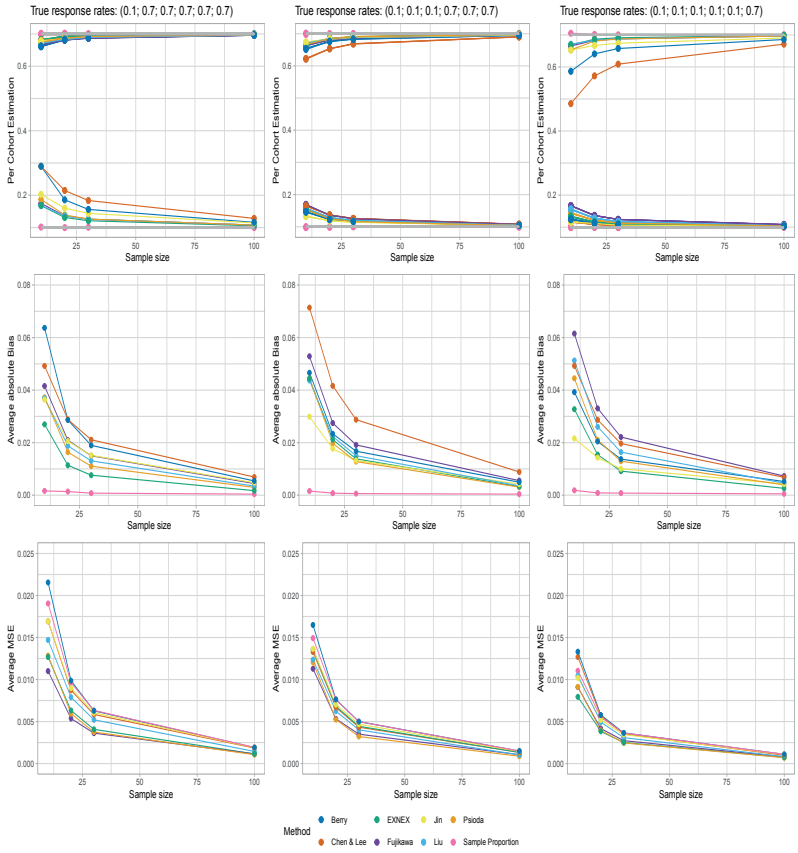


Figure 4.12: *Estimates on scenarios 2.D.1 to 2.D.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

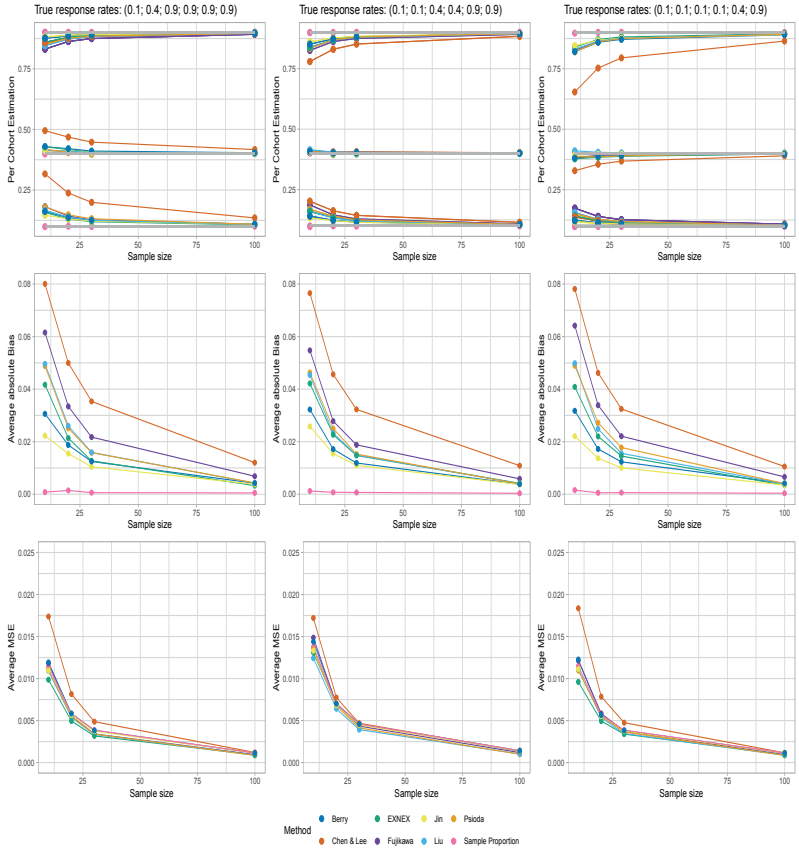


Figure 4.13: *Estimates on scenarios 3.B.1 to 3.B.3, in 4 sample size points, 10, 20, 30, 100, in the 1st row. In the second row, is the average absolute Bias of the respective scenarios and the third row is the average of MSE.*

5

Evaluating Basket Trial Methodology in Oncology: A Comparative Analysis Using the DRUP study

Authors: Antonios Daletzakis, Rutger van den Bor, Kit CB Roes, Vincent van der Noort

Original title: Evaluating Basket Trial Methodology in Oncology: A Comparative Analysis Using the DRUP study

Under review

Abstract

In the era of precision medicine, basket trials have emerged as a core methodology for evaluating the activity of treatments across different patient populations with shared molecular characteristics. This study investigates the application of various Bayesian methods in the context of the DRUP (Drug Rediscovery Protocol) trial, which employs a Simon's two-stage design (STS) for each sub-study within a basket. These methods have in common that they allow for borrowing of information across cohorts. Focusing on estimation of the response rates in each of the baskets, simulations were used to determine optimal values for each method's required parameter settings, as well as to compare the performance of the different estimation procedures. The results underscore the importance of a priori selecting optimal tuning parameters tailored to specific trial conditions to enhance the reliability of response rate estimates. The different simulation settings suggest that the EXNEX method may offer enhanced robustness compared to the other evaluated methods. This study provides practical recommendations for applying these methods in basket trials, aiming to improve the accuracy of estimation of treatment efficacy in basket trials in oncology.

5.1 Introduction

In the evolving landscape of precision medicine, master protocol studies have emerged as a pivotal methodology, particularly in (non-randomised) exploratory phase I/II research. [12] [11] Basket trials that we address in this paper involve multiple substudies of the same drug, each with a single-arm trial design under a unified treatment protocol. The basket trials target patients with varying tumor histologies that share a common molecular profile, related to the mechanism of action of the drug. This innovative approach allows for the simultaneous evaluation of treatment effects across different patient populations, optimizing resource utilization and accelerating the clinical development process.

Basket trials present considerable methodological challenges, in terms of design and statistical analysis. A single or two-stage design is usually proposed for substudies in this methodology. Kasim et. al [17] conducted an extensive literature review of ongoing, completed and terminated basket trials, identifying that out of 79 trials, 41 are using two-stage designs or fully sequential designs. Simon's two-stage design (STS)[16], was used in 30 basket trials. However, Kasim et. al [17] highlights that the lack of standardized statistical tools and clear guidance presents a challenge for consistent implementation and reproducibility of such trials.

The STS design is thus a popular design choice in exploratory trials with the ability to balance between preventing exposure to ineffective treatments and power while minimizing the required sample size. Although alternatives for the STS design have been proposed (by Zhou et. al [65], Wu et. al [66] and Jing et. al [67]) the STS design remains widely used practice. The Drug Rediscovery Protocol (DRUP)[43] trial is an ongoing prospective multi-drug and pan-cancer trial. Patients with progression of an advanced or metastatic solid tumour, multiple myeloma or B-cell non-Hodgkin lymphoma who lack standard treatment options, are eligible for this trial. DRUP consists of a collection of baskets of substudies designed with a STS design. Baskets are defined based on the combination of a molecular alteration shared by the patients and a drug given in the study. Some baskets are subdivided into different substudies based on histology, some consist of a single, tumor-agnostic sub-study. All drugs studied in the DRUP trial have already been approved for other tumor types than the ones studied here, hence the name drug-rediscovery. The primary purpose of the STS design is to make a go/no-go decision on whether the drug studied seems promising or not in the given patient group. The DRUP trial is designed with STS design assuming a Null hypothesis that the clinical benefit rate of 0.1 is unacceptable and an alternative hypothesis of 0.3 clinical benefit

rate is a desired effect size. The "type I error" alpha, used across all cohorts, is 0.078 at a power of 0.85. According to the admissible design as proposed by Jung et. al [68] the trial proceeds to a second stage if at least 1 response is observed among the first 8 patients, with a total of 24 patients required to confirm treatment is not futile if at least five responses are recorded. Clinical benefit in this context is defined as a complete response (CR), partial response (PR), or stable disease lasting at least 16 weeks.

The estimation of response (or clinical benefit) rates within a two-stage design introduces additional statistical complexities. Porcher et. al [30] suggested how to obtain proper inference on the response rate in a STS design, and noted that the sample proportion is a biased estimate (underestimation of the true response rate) [21] for sequential designs. Alternative estimators were proposed that reduce the bias or provide unbiased estimates. These include the uniformly minimum-variance unbiased estimator (UMVUE) for the multi-stage approach proposed by Jung et. al [22]. Many basket trials aim to estimate the response rate by either pooling across all cohorts or analyzing each cohort independently [69]. Each strategy has its drawbacks, the pooled strategy has the risk to inflate the type I error in case of large heterogeneity of treatment effects, while the independent analysis in small heterogeneity scenarios lacks statistical power compared to alternative methods. Bayesian methods allow for estimation of the basket-specific response rates, while making use of results from other baskets. Such approaches may improve the performance of the estimators, particularly when sample sizes are limited. [17], [14]. A methodological review for the estimation of the response rate in a basket trial using single stage designs for sub-studies can be found in our previous work [70].

This paper aims to evaluate and compare various estimation procedures for the response rates that allow for borrowing, from an application perspective, using the DRUP trial as example. The methods require specification of prior distributions and tuning parameters, which may affect the performance of the estimators. As it is important to pre-specify analysis methods in full at the design stage, we use the design and characteristics of the DRUP substudies in different baskets to illustrate an approach to set these parameters a priori, as well as compare the different methods. A similar approach as Sauer et. al [71] and Baumann et. al [72] is followed, that suggest methods to select optimal tuning parameters through the maximisation of the expected number of correct decisions (ECD) under a collection of various scenarios. As our focus is on estimation rather than decision making, we use average Root Means Square Error (aRMSE) based metrics (the min-mean aRMSE and the min-max aRMSE) as utility functions, evaluated over a range of selected scenarios. The range of possible scenarios given the design choices of the DRUP study is selected as basis to rationally select prior(s) and tuning parameters for each method and evaluate the methods. The overall approach provides the basis for suggestions to the researches for the optimal application of the methods.

Section 5.2 details the basket trial examples selected from the DRUP study and presents the response rate estimation based on the parameters as suggested in the original papers. Section 5.3 explains the simulation scenario sets and discusses parameter optimization given the different utility function criteria. Results of both parameter choice and the resulting estimation for the DRUP baskets are presented in section 5.4 and the Appendix A. The discussion section 5.5 addresses the strengths and limitations and offers suggestions for application of our approach in design and analysis of basket trials with the aim to provide reliable estimation.

5.2 Methodology

DRUP trial example and initial estimation

Three different examples of actual basket trials under the DRUP study are used in this paper. The DRUP trial is an ongoing trial and not all of the baskets used in this paper are published yet, so the histologies and the genomic aberration types are masked. In the DRUP platform trial a large number of baskets with multiple substudies have already opened, of which only a small part has been completed. In table 5.1 the actual results (response) of the three basket trial examples from the DRUP study are presented. Basket trial 1 includes 4 cohorts (substudies), of which Cohort 1D did not complete the first stage yet (8 patients) and the remaining cohorts are ongoing in stage 2. Cohorts 1A and 1C have already reached the designs target of at least 5 responses to claim that the treatment is active for this genomic aberrations. Basket trial 2 has already 2 completed and successful cohorts and 2 cohorts recruiting patients in the second stage. Basket trial 3 is the one with the most completed cohorts so far. Cohorts 3A, 3B and 3C have been completed, with the cohort 3A having one patient less than the required design, but the number of responses is more than needed to reach success in this STS design. Cohort 3C has one more patient than the design requires, because during the accrual two patients where eligible to take the treatment at that time. Cohort 3D is still recruiting in the second stage of the design.

Basket trial 1 (Lenvatinib)		
	Clinical benefit	Total Patient
Cohort 1A	6	16
Cohort 1B	3	14
Cohort 1C	8	11
Cohort 1D	3	5
Basket trial 2 (Trastuzumab)		
Cohort 2A	9	24
Cohort 2B	11	24
Cohort 2C	4	19
Cohort 2D	3	8
Basket trial 3 (Olaparib)		
Cohort 3A	14	23
Cohort 3B	10	24
Cohort 3C	8	25
Cohort 3D	3	17

Table 5.1: DRUP basket trials and the respective cohorts masked

We examine how the selected estimation methods can be used for estimation of the response rates, given the STS design in each cohort, including the scenario that the targeted sample size is not yet reached. It is important for practical applications to keep in mind that it might not be feasible to complete the pre-specified sample size for all substudies. The natural option for estimation is the sample proportion: the observed number of responses divided by the total number of patients. This is, however, biased for most cohorts. The alternative Jung estimator for the analysis of independent cohort level response rate following the STS design might not be appropriate in this case, since the trial sample size is not completed for a number of cohorts. Berry et. al [54], EXNEX [55], Psioda et. al [57] and Fujikawa et. al [58], described in more detail below, are Bayesian methods to estimate the response rates in each basket trial. To conduct an analysis using these methods, parameters need to be specified. The authors proposed tuning parameters and prior distribution parameters. Following our previous work [70], if flat priors are applied to the example data from DRUP, results are as displayed in Table 5.2 below. As can be seen from Table 5.2,

the response rates estimates presented can vary substantially depending on the estimator used.

Basket trial 1 (Lenvatinib)						
	x/n	Sample prop	Psioda	Fujikawa	Berry	EXNEX
Cohort 1A	6/16	0.375	0.385	0.369	0.416	0.391
Cohort 1B	3/14	0.214	0.282	0.310	0.363	0.261
Cohort 1C	8/11	0.727	0.664	0.655	0.533	0.673
Cohort 1D	3/5	0.6	0.548	0.542	0.480	0.543
Basket trial 2 (Trastuzumab)						
Cohort 2A	9/24	0.375	0.379	0.386	0.362	0.371
Cohort 2B	11/24	0.4583	0.430	0.422	0.374	0.425
Cohort 2C	4/19	0.2105	0.277	0.315	0.340	0.264
Cohort 2D	3/8	0.375	0.381	0.382	0.361	0.371
Basket trial 3 (Olaparib)						
Cohort 3A	14/23	0.609	0.568	0.6	0.473	0.570
Cohort 3B	10/24	0.417	0.413	0.382	0.401	0.407
Cohort 3C	8/25	0.320	0.337	0.339	0.364	0.336
Cohort 3D	3/17	0.177	0.248	0.269	0.320	0.225

Table 5.2: DRUP basket trials and the respective cohorts masked. Numbers represent the estimate of the clinical benefit rate obtained by each method using a flat prior

The methodological question that arises is whether the estimates presented in Table 5.2 are optimal for the given design or if a different (a priori) choice of parameter values can yield better estimates. We investigate this by optimizing parameter selection based on the minimization of the Root Mean Square Error (RMSE) under a suitable collection of scenarios. This is done through the simulation study described below. It is noted that such a simulation study can be performed at the design stage, generally allowing for pre-specification of the estimation method.

The four Bayesian methods evaluated are summarised as follows.

Estimator based on Berry et al.[54]

The Berry method employs a Bayesian hierarchical model (BHM) to estimate the response rates in different cohorts, assuming exchangeability between cohorts, which allows for information sharing.

The observed responses y_j for each cohort j follow a binomial distribution:

$$y_i \sim \text{Binomial}(n_j, \pi_j)$$

The logit-transformed response rates $\theta_j := \text{logit}(\pi_j)$ are assumed to be normally distributed:

$$\theta_j \sim \mathcal{N}(\mu, \sigma^2)$$

The hyperparameters for the priors are:

$$\mu \sim \mathcal{N}(0, 100), \sigma^2 \sim \text{IG}(0.0005, 0.00005)$$

The posterior distribution is obtained by updating the priors with the observed data using Bayes' theorem. Information is borrowed between cohorts through the hierarchical structure, assuming that the response rates are exchangeable.

Estimator based on Neuenschwander et al.[55]

The EXNEX (Exchangeable-Non-Exchangeable) method, proposed by Neuenschwander et al.[55] , combines exchangeable and non-exchangeable components to model response rates in different cohorts, providing a flexible framework for information borrowing.

The observed responses y_j for each cohort j follow a binomial distribution:

$$y_j \sim \text{Binomial}(n_j, \pi_j)$$

The logit-transformed response rates $\theta_j := \text{logit}(\pi_j)$ are modelled using a mixture of exchangeable and non-exchangeable components:

$$\theta_j|\mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2) \text{ with probability } w$$

$$\theta_j \sim \mathcal{N}(m_i, v_i) \text{ with probability } 1 - w$$

The hyper-parameters for the priors are:

$$\mu \sim \mathcal{N}(0, 3), \tau^2 \sim \text{half-normal}(1)$$

The mixture weights w determine the extent of borrowing between cohorts. It is a fixed value that needs to be set beforehand. If w is high, more information is borrowed from other cohorts.

The EXNEX method uses Markov Chain Monte Carlo (MCMC) methods to estimate the posterior distributions of the response rates, accounting for both exchangeable and non-exchangeable components. For the purposes of this paper the 'bhmbasket' R package is used for the calculation of the estimate.

Estimator based on Psioda et al.[57]

The Psioda method employs a Bayesian Model Averaging (BMA) approach to enhance the estimation of cohort-specific response rates in basket trials. This method combines information from multiple cohorts by averaging over all possible models that describe the relationships between cohorts.

The observed responses y_j for each cohort j follow a binomial distribution:

$$y_i \sim \text{Binomial}(n_j, \pi_j)$$

where n_j is the number of subjects in cohort j and π_j is the response rate for cohort j. The prior distribution for the response rate π_j is given by a Beta distribution:

$$\pi_i \sim \text{Beta}(a_j, b_j)$$

The posterior distribution of π_j is obtained by updating the prior with the observed data:

$$\pi_i|y_i \sim \text{Beta}(a_j + y_j, b_j + n_j - y_j)$$

Model averaging is performed by computing the posterior distribution for each possible model and then averaging these distributions, weighted by the posterior model probabilities.

Information borrowing is achieved by assigning probabilities to all possible models that include different combinations of cohorts. The posterior distributions of the response rates are weighted averages of the model-specific posterior distributions, allowing for adaptive borrowing based on the similarity of response rates across cohorts. Estimates were obtained using the function 'bma' (version 0.1.2) in the R package 'bmabasket'.

Estimator based on Fujikawa et al. [58]

The Fujikawa method is a Bayesian approach that borrows information across cohorts based on the similarity of their response rates. This method uses a similarity measure to determine the extent of borrowing, allowing for more accurate estimation of response rates when cohorts are similar.

The observed responses y_j for each cohort j follow a binomial distribution:

$$y_j \sim \text{Binomial}(n_j, \pi_j)$$

The prior distribution for the response rate π_j is a Beta distribution:

$$\pi_j \sim \text{Beta}(a_j, b_j)$$

The posterior distribution is updated based on the observed data:

$$\pi_j | y_j \sim \text{Beta}(a_j + y_j, b_j + n_j - y_j)$$

The method employs similarity measures such as the Jensen-Shannon (JS) divergence or the Kullback-Leibler (KL) distance to quantify the similarity between cohorts. The similarity measure w_{jh} between cohorts j and h is defined as:

$$w_{jh} = 1 - \text{JS}(P \parallel Q)$$

Borrowing is allowed if the similarity measure exceeds a threshold τ. The updated prior for π_i is a weighted sum of the Beta distributions from similar cohorts:

$$\pi_j \sim \text{Beta}\left(\sum_{h=1}^{K} I(w_{ih} > \tau) w_{ih}^{\epsilon} a'_h, \sum_{h=1}^{K} I(w_{ih} > \tau) w_{ih}^{\epsilon} b'_h\right)$$

The method adaptively borrows information based on the similarity between cohorts.

5.3 Simulation study for parameter selection

Response data for four cohorts of patients are simulated based on a binomial distribution with specified probabilities p. The simulations were designed to reflect the DRUP trial and incorporate the STS-based decision criteria for each cohort.

We simulated with probabilities $p = 0.1, p = 0.3$, and $p = 0.6$ to generate patient response data. Each simulation scenario was repeated $S = 10,000$ times.

Scenario set 1 and 2 followed completely the STS design, so the possible sample sizes could be either 8, if stopping in the first stage is decided, or 24 if the first stage is completed successfully. The Scenario set "reduced" addresses

that in practice the total STS sample size may not be reached, due to various reasons (e.g., recruitment). It follows the same response rate scenarios as the Scenario set 1, table 5.3. For "Scenario reduced", the first stage is assumed to be completed and if it is possible to proceed in the second stage, the trial is assumed to stop after a random number of patients in the second stage. This randomness is implemented by choosing the second stage sample size by a uniform distribution, where the minimum number of patients is 0 and the maximum is 16.

The following scenarios were thus simulated:

1. **Scenario Set 1**: Probabilities p were set to low (0.1), mid (0.3), and high (0.6) values for different combinations across the four cohorts.
2. **Scenario Set 2**: To test the sensitivity of the tuning parameters under extreme conditions, probabilities were set to 0.4, 0.7, and 0.9.
3. **Scenario Set reduced**: To test the sensitivity of the tuning parameters under a realistic setting, where the sample size is smaller than in Scenario Set 1, probabilities were set to 0.1, 0.3, and 0.6.

The specific combinations of response probabilities across the cohorts in each scenario set are detailed in Table 5.3 and Table 5.7 respectively.

	Cohort A	Cohort B	Cohort C	Cohort D
a	0.1	0.1	0.1	0.1
b	0.1	0.1	0.1	0.3
c	0.1	0.1	0.1	0.6
d	0.1	0.1	0.3	0.3
e	0.1	0.1	0.3	0.6
f	0.1	0.1	0.6	0.6
g	0.1	0.3	0.3	0.3
h	0.1	0.3	0.3	0.6
i	0.1	0.3	0.6	0.6
j	0.1	0.6	0.6	0.6
k	0.3	0.3	0.3	0.3
l	0.3	0.3	0.3	0.6
m	0.3	0.3	0.6	0.6
n	0.3	0.6	0.6	0.6
o	0.6	0.6	0.6	0.6

Table 5.3: Scenarios selected in the simulation studies for Scenario Set 1 and Scenario Set reduced. Numbers represent the true response probability in each cohort in each scenario

The **RMSE** for each cohort j is defined as the square root of the mean of the squared differences between the estimated and true values. It is calculated as:

$$\text{RMSE}_j = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(\hat{p}_{ij} - p_j)^2}$$

where $\hat{p}_{ij}$ is the estimated value in simulation i and p_j is the true value for each cohort j. The average RMSE (aRMSE) across all cohorts is then calculated as:

$$\text{aRMSE} = \frac{1}{4}\sum_{j=1}^{4}\text{RMSE}_j$$

In order to optimize the selection of the different parameter choices in each method, two strategies are followed:

1. **minimization of the Mean of the aRMSE**: An optimal combinations of parameter values is selected by minimizing the mean aRMSE across all scenarios.

2. **minimization of the Maximum of the aRMSE**: An optimal combination of parameter values is selected by minimizing the maximum aRMSE. It thus controls that the selected parameter set does not have an exceptionally high aRMSE in any scenario.

For the notation of the paper we will use the following utility functions:

1. $\hat{U}_{\textbf{min-mean aRMSE}}$: Indicating the minimization of the Mean of the aRMSE across all scenarios.

2. $\hat{U}_{\textbf{min-max aRMSE}}$: Indicating the minimization of the Maximum of the aRMSE across all scenarios.

By exploring a comprehensive parameter space, we aim to identify the parameter settings for each Bayesian method that ensure reliable and accurate response rate estimates in the context of basket trials using the STS design.

5.3.1 Parameter Space for Simulation

Each method has a number of parameters and priors to specify. In our previous research work [70], we used flat-uninformative priors and parameters as suggested by the authors. In this paper, we explore a variety of parameter combinations within the parameter space for each method.

For Berry's method, the following parameter space is explored:

$$\begin{aligned}
\text{Target mean:} \quad & [0.1, 0.5, 0.9] \\
\text{Variance parameter:} \quad & \tau \sim \text{dgamma}(\tau_{disp}, \tau_{disp}/100), \quad \tau_{disp} = [0.00005, 0.0005, 0.005] \\
\text{Hyper-parameters of the mean:} \quad & \begin{cases} \mu = [-2, -1, 0, 1, 2] \\ \text{var} = [0.001, 0.01, 0.1, 1] \end{cases}
\end{aligned}$$

*

For the EXNEX method the following parameter space is explored, including different values for the weight between the EX and NEX parts:

$$\begin{aligned}
w &= [0.1, 0.3, 0.5, 0.7, 0.9] \\
\tau^2 &= [0.1, 0.75, 1, 1.25] \\
\text{EX mean:} &= [0.1, 0.5] \\
\text{NEX mean:} &= [0.1, 0.5]
\end{aligned}$$

For the Psioda method [57], three different parameters need to be set: the mean of the prior, the dispersion parameter, and the prior distribution applied to the possible model choice. The explored parameter space is as follows:

$$\begin{aligned}
\text{Prior mean:} \quad & \mu = [0.1, 0.3, 0.5, 0.7, 0.9] \\
\text{Dispersion parameter:} \quad & \phi = [0.5, 1, 2, 3, 5, 8, 10] \\
\text{Prior model parameter:} \quad & pmp = [-10, -8, -6, -4, -3, -2, -1, 0, 1, 2, 3, 4, 6, 8, 10]
\end{aligned}$$

*The number of examined tuning parameters and prior parameters was limited because of the computational time required to run all the possible scenarios. The Psioda method was fastest so the parameter space was extended.

The pmp parameter assigns relative weights to the independent (heterogeneity between cohorts) model (positive value) or to the homogeneous model (negative value). Although Psioda initially proposed using zero or positive values for the pmp parameter, we extend this by also exploring negative values. This allows us to examine scenarios favouring homogeneous models, thus providing a comprehensive evaluation of the method's performance under a wider range of cohort similarity assumptions.

For the Fujikawa method, the following parameter space is explored:

$$\begin{aligned} \alpha &= [1] \\ \beta &= [1, 2.33333] \\ \tau &= [0, 0.1, 0.3, 0.5, 0.7, 0.9, 1] \\ \epsilon &= [0.5, 1, 1.5, 2, 2.5, 3, 4, 6] \end{aligned}$$

5.4 Results

In Fig. 5.1, the mean aRMSE of all scenarios for the four examined methods are presented. The overall graphical representation of the results is based on a colour palette (heat-map) where green indicates the lowest mean aRMSE, black the highest, and white, yellow and red represent intermediate values. In all heat-map figures, the mean or max aRMSE value is presented using the same colour limits, so the graphs can be directly compared.

Berry's method exhibited relatively uniform performance across parameter combinations, showing only minor variations in both mean and max aRMSE (fig 5.2,5.1). The tuning parameters influenced mean aRMSE slightly, but max aRMSE remained largely unaffected across parameter settings. In general, the method showed limited sensitivity to prior specification, with small differences in performance between parameter combinations. The "target mean" parameter, reflecting the expected response rate, had an optimal value of 0.5 in most cases. However, differences between the two utility functions, min-mean and min-max aRMSE were observed in other prior settings. For example, the optimal prior mean for the normal distribution was -2 for min-mean aRMSE in scenario set 1, while for min-max aRMSE the values shifted to 1 in scenario set 1 and 2 in the Reduced set patient scenarios. The optimal prior variance remained fixed at 1 across all cases, despite the authors recommending a smaller value (0.01), indicating that a higher dispersion, which does not allow the prior to spread more around its mean, seems to improve the estimation process. Additionally, the choice of the parameter of inverse gamma distribution for the variance had minimal effect on the results. These findings suggest that Berry's method is relatively stable under different parametrizations, on average. Finally, although bias was generally low across all settings, an expected trade-off was observed: parameter sets with the lowest mean absolute bias tended to produce higher mean aRMSE values.

The EXNEX method, which combines exchangeable (EX) and non-exchangeable (NEX) components, demonstrated consistent performance in terms of mean and max aRMSE across the different tuning parameters. Optimal aRMSE values were observed when both the EX and NEX prior means were set to 0.5, with minimization of the aRMSE occurring at a high weight parameter ($w = 0.9$), favouring more borrowing of information across baskets. In both mean aRMSE and max aRMSE, fig 5.2,5.1, it was observed that as the weight (w) increased in favour of the EX component, thus allowing more borrowing, the aRMSE values decreased. However, an exception was noted at lower values of the prior variance parameter (tau.HN = 0.1), where increasing the weight (w) actually led to increased aRMSE values. Other EX and NEX prior mean parameter values exhibited the same general pattern as the optimal 0.5 setting, but resulted in slightly higher mean and max aRMSE values. The optimal values for the tau.HN parameter ranged from 1 for the min-mean aRMSE scenario set 1 to 1.25 for both the min-max aRMSE and the reduced patient scenarios, aligning closely with the authors' suggested value of 1. Optimal parameters remained similar

under the "Reduced set" scenarios. Additionally, mean absolute bias was consistently low across parameter sets, with the smallest mean aRMSE observed when a moderate level of bias was allowed.

Psioda's method was influenced by the selection of the parameters, the prior mean (μ_0), dispersion (ϕ_0), and the prior model parameter (pmp_0). In general, the method followed consistent patterns across parameter combinations, though the performance differed across utility functions and settings. Regarding the prior mean (μ_0), a value of 0.3 resulted in the lowest mean aRMSE, while 0.5 was optimal for minimizing max aRMSE. These values aligned closely with the average true response rate across the scenario set. In scenario set 1 the average true response rate is 0.33 and in scenario set 2 is 0.7. Similarly, the prior model parameter (pmp_0) had a clear effect on aRMSE values. For min-mean aRMSE, lower values such as $pmp_0 = 0$ or slightly above led to better performance, favouring the uninformative prior model. In contrast, for min-max aRMSE, higher values were preferred ($pmp_0 = 4$ in scenario set 1 and $pmp_0 = 6$ in the reduced patient scenario), indicating better performance when assigning more weight to models with more parameters, like the independent model. While the authors recommended a weakly informative prior with $\phi_0 = 1$, our simulations indicated that higher dispersion values improved estimation performance. The smaller aRMSE values in parameter ϕ_0 ranged from 1 to 5, with $\phi_0 = 5$ providing the best results for most settings, and $\phi_0 = 3$ performing best in the reduced sample size scenario under min-max aRMSE. Psioda's method emphasizes the importance of careful parameter tuning. While optimal settings differ slightly between the two utility functions, multiple parameter combinations can lead to good estimation performance. A trade-off was observed between aRMSE and bias, where parameter sets yielding lower aRMSE did not always minimize mean absolute bias. This underlines the need to balance precision and accuracy when selecting prior settings for Psioda's method.

The Fujikawa method facilitates information borrowing across baskets by adjusting two key parameters: τ, which determines whether borrowing is permitted, and ϵ, which regulates the extent of borrowing. While Fujikawa et al. caution against using $\epsilon > 2$ to avoid underestimating similarity between baskets, our simulation results suggest otherwise. We find that the performance — measured by both mean and max aRMSE — varies notably with parameter choices. Specifically, combinations of small τ (≤ 0.5) and small ϵ (≤ 2) lead to higher error. In contrast, better performance is achieved when either $\tau > 0.5$ or $\epsilon > 2$, regardless of the other parameter's value. Interestingly, our findings consistently indicate that setting $\tau = 0$ — allowing borrowing in all cases — yields optimal outcomes across utility functions. Moreover, although larger ϵ values (e.g., $\epsilon = 6$) reduce the degree of borrowing, the differences in estimation accuracy and aRMSE compared to moderate values (e.g., $\epsilon = 4$) remain minimal. The beta prior parameters seem to play an important role in the mean and max aRMSE. The optimal value of $\alpha = 1$ and $\beta = 2.333$ is observed for the Reduced set and the full patient scenarios but in scenario set 2,the optimal values are $\alpha = 1$ and $\beta = 1$ as suggested by the authors as well.

Method	Parameter	$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-mean aRMSE reduced}}$	$\hat{U}_{\text{min-max aRMSE}}$	$\hat{U}_{\text{min-max aRMSE reduced}}$	Uninformative
Fujikawa	τ	0	0	0	0	0.5
	ϵ	4	6	4	6	2
	α	1	1	1	1	1
	β	2.333	2.333	2.333	2.333	1
Psioda	ϕ_0	5	5	5	3	2
	pmp_0	0	0	4	6	1
	μ_0	0.3	0.3	0.5	0.5	0.5
Berry	$mean$	0.5	0.5	0.5	0.5	0.5
	$prior_{\mu.mean}$	-2	-2	1	2	0
	$prior_{\mu.var}$	1	1	1	1	0.01
	τ_{disp}	0.005	0.005	0.005	0.005	0.0005
EXNEX	$ex.p.w$	0.5	0.5	0.5	0.5	0.5
	$nex.p.w$	0.5	0.5	0.5	0.5	0.5
	w	0.9	0.9	0.9	0.9	0.5
	$tau.HN.scale$	1	1.25	1.25	1.25	1

Table 5.4: Parameters optimal values on the different utility functions

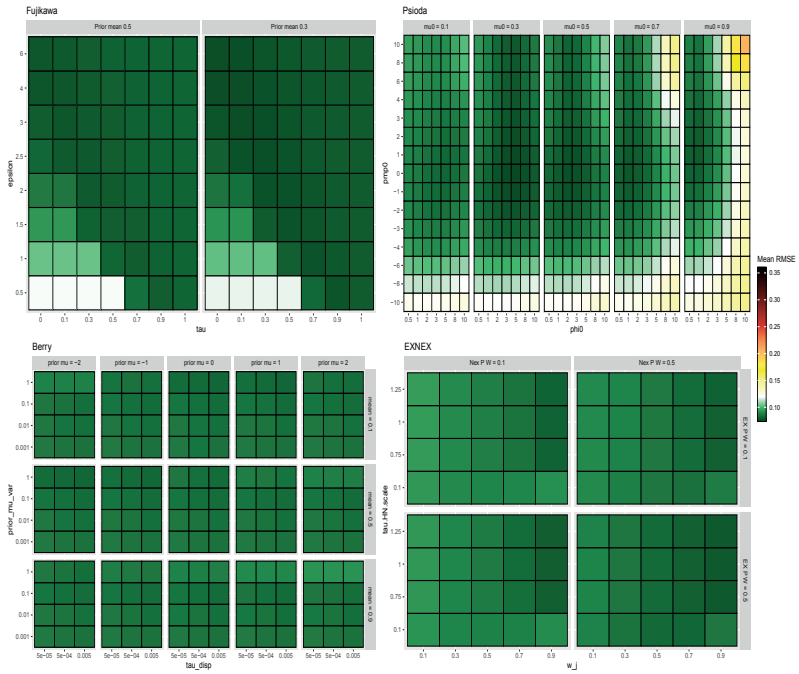


Figure 5.1: Results of the **scenario set 1**. A heat-map figure of the **Mean aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

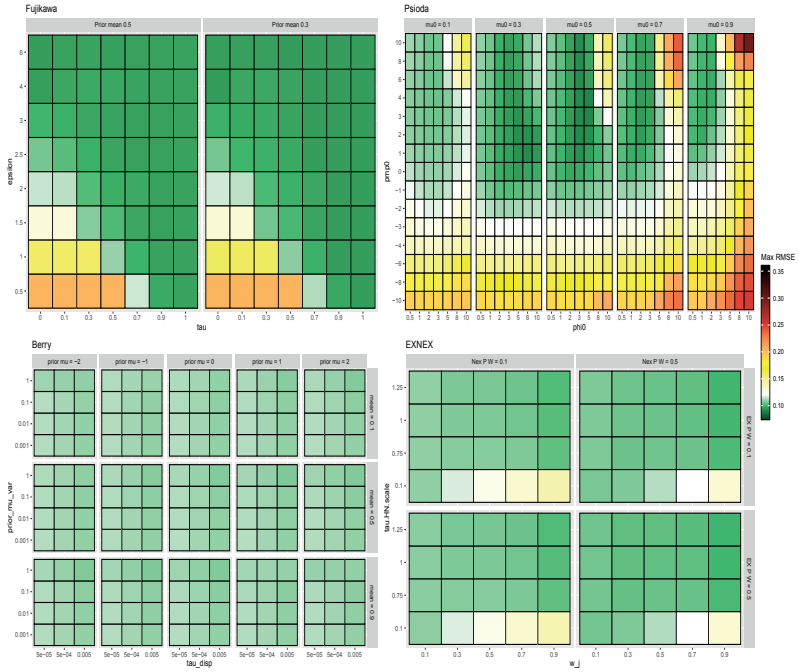


Figure 5.2: Results of the **scenario set 1**. A heat-map figure of the **Max aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

5.4.1 Sensitivity analysis

The Berry method is slightly sensitive in the target mean parameter. In scenario set 1 the 0.1 and 0.5 are showing the best aRMSE values in both mean or max functions, and in scenario set 2 the 0.5 and 0.9 are having smaller error values. Although the observed differences in performance are relatively minor, acknowledging these variations is important for a comprehensive understanding.

The EXNEX method does not show sensitivity in the patterns of aRMSE across the two scenario sets, or even when a reduced number of patients is used, indicating that its parameter choices are robust to variations in the true response rates.

For the Psioda method, sensitivity to the prior mean choice is evident. In scenario set 1, the smallest mean aRMSE values occur when the prior mean was set to 0.3, matching the average response probability in that scenario set. Similarly, in scenario set 2, a prior mean of 0.7 yields the lowest mean aRMSE values. These results are an indication that the Psioda method's mean is sensitive to the underlying scenario conditions. This effect also exists in the Reduced set scenario (Fig 5.1,5.2,5.5 for the mean and fig 5.9 for the max aRMSE).

For the Fujikawa method, similar sensitivity to the prior mean is observed, between the two scenario sets. Mean aRMSE is more affected by prior choices than max aRMSE, suggesting that this method is relatively robust to extreme errors but more sensitive to average-case performance. In scenario set 1, where the average true response rate is 0.3, a prior mean of 0.3 gives the smallest aRMSE values. In scenario set 2, where the average true response rate is 0.7, a prior mean of 0.5 gives a more accurate aRMSE. Scenario set reduced suggested the optimal $\epsilon = 6$, which is higher value than in Scenario set 1, $\epsilon = 4$. This choice does not greatly impact the results, even the difference in aRMSE is really small in every scenario set.

The Bayesian methods consistently outperformed the sample proportion in terms of mean and max aRMSE. In scenario set 1, under min-mean aRMSE, the sample proportion estimator had a mean aRMSE of 0.099, while Bayesian methods showed lower values: Fujikawa (0.077), Psioda (0.076), EXNEX (0.080), and Berry (0.083) (Table 5.5). Although Bayesian methods introduced slightly higher mean absolute bias compared to the sample proportion estimator, ranging from 0.022 to 0.034 versus 0.013, they improved overall estimation precision, clearly reflected by their lower aRMSE values. Similarly, under the min-max aRMSE scenario, Bayesian methods continued to present lower aRMSE values than simpler estimators. For example, the max aRMSE for the sample proportion estimator was higher (0.113) compared to Bayesian approaches like Fujikawa (0.095), Psioda (0.090), EXNEX (0.101), and Berry (0.107). Even in reduced patient scenarios, Bayesian methods maintained their advantage: while the sample proportion estimator reached a max aRMSE of 0.132, Bayesian methods such as Fujikawa (0.113) and Psioda (0.107) provided considerably lower values. A further comparison between optimized and uninformative priors reveals that tuning can improve performance. For example, in scenario set 1 under the min-mean aRMSE utility, Fujikawa's method had a mean aRMSE of 0.077 with optimized priors versus 0.080 with uninformative priors, and corresponding mean absolute biases of 0.034 and 0.022, respectively. Psioda showed a similar pattern, with mean aRMSE increasing from 0.076 to 0.078 and mean absolute bias decreasing from 0.031 to 0.017 when moving from optimized to uninformative priors. While uninformative priors sometimes led to lower bias, the optimization resulted in better aRMSE values.

5.4.2 DRUP trial results with selected parameters

The simulation study identified optimal parameter choices for each utility function used to evaluate the methods: min-mean aRMSE and min-max aRMSE. These optimal parameters were derived based on the specific design of the DRUP study's Simon's two-stage design (STS), but using two possible settings. Strictly follow the design and simulate full cohorts, or simulate a reduced number of patients, given a random stop after completing the first stage. Table 5.4 presents the results of these optimizations, while Table 5.6 provides the corresponding response rate estimates for each basket.

In table 5.5, the mean aRMSE and the max aRMSE for the two different optimal utility functions and settings and the uninformative parameters are presented. The Bayesian methods have a smaller mean aRMSE and max aRMSE compared to the sample proportion in almost all cases. The results across the Bayesian methods appear similar at the parameter settings optimised with the above approach. The results between the Bayesian methods are comparable in terms of aRMSE, but the EXNEX method has smaller mean bias at the selected parameters for both utility functions.

scenario set 1						
	$\hat{U}_{\text{min-mean aRMSE}}$		$\hat{U}_{\text{min-max aRMSE}}$		Uninformative	
Method	mean aRMSE (mean abias)	max aRMSE (max abias)	mean aRMSE (mean abias)	max aRMSE (max abias)	mean aRMSE (mean abias)	max aRMSE (max abias)
Sample proportion	0.099 (0.013)	0.113 (0.028)	-	-	-	-
Fujikawa	0.077 (0.034)	0.095 (0.043)	0.077 (0.034)	0.095 (0.043)	0.08 (0.022)	0.099 (0.037)
Psioda	0.076 (0.031)	0.095 (0.039)	0.088 (0.044)	0.09 (0.080)	0.078 (0.017)	0.095 (0.024)
EXNEX	0.08 (0.022)	0.101 (0.037)	0.080 (0.02)	0.101 (0.033)	0.084 (0.01)	0.104 (0.019)
Berry	0.083 (0.031)	0.109 (0.055)	0.086 (0.031)	0.107 (0.055)	0.086 (0.031)	0.111 (0.057)
scenario set reduced						
	$\hat{U}_{\text{min-mean aRMSE}}$		$\hat{U}_{\text{min-max aRMSE}}$		Uninformative	
	mean aRMSE (mean abias)	max aRMSE (max abias)	mean aRMSE (mean abias)	max aRMSE (max abias)	mean aRMSE (mean abias)	max aRMSE (max abias)
Sample proportion	0.117 (0.01)	0.132 (0.019)	-	-	-	-
Fujikawa	0.093 (0.039)	0.113 (0.053)	0.093 (0.039)	0.113 (0.053)	0.099 (0.034)	0.118 (0.051)
Psioda	0.092 (0.044)	0.11 (0.057)	0.102 (0.04)	0.107 (0.072)	0.096 (0.026)	0.113 (0.034)
EXNEX	0.097 (0.029)	0.118 (0.045)	0.097 (0.029)	0.118 (0.045)	0.101 (0.014)	0.122 (0.025)
Berry	0.1 (0.044)	0.126 (0.069)	0.107 (0.034)	0.108 (0.061)	0.103 (0.042)	0.111 (0.074)

Table 5.5: aRMSE and abias values on the different utility functions for scenario set 1 and reduced number of patients

Table 5.6 summarizes the response rate estimates for each method, based on the optimal parameter configurations identified through the simulation study. Across all methods, there was a clear tendency for the estimates to shrink toward the mean, regardless of parameter choice. The estimated response rates of every method are lower than the sample proportions in Cohort 1C, where 8 out of 11 patients responded to treatment. The methods provide estimates that are larger than the sample proportion for Cohort 3D. This discrepancy highlights the challenge of balancing borrowing across baskets while maintaining accuracy in individual cohort estimates.

			Basket 1 (Lenvatinib)											
Cohort	x/n	Sample prop	Psioda			Fujikawa			Berry			EXNEX		
			$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-max aRMSE}}$	Uninf	$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-max aRMSE}}$	Uninf	$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-max aRMSE}}$	Uninf	$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-max aRMSE}}$	Uninf
Cohort 1A	6/16	0.375	0.377 (0.377)	0.404 (0.395)	0.385	0.365 (0.361)	0.365 (0.361)	0.369	0.385 (0.385)	0.433 (0.439)	0.416	0.403 (0.401)	0.401 (0.401)	0.391
Cohort 1B	3/14	0.214	0.284 (0.284)	0.291 (0.265)	0.282	0.296 (0.275)	0.296 (0.275)	0.310	0.321 (0.321)	0.370 (0.361)	0.363	0.304 (0.297)	0.297 (0.297)	0.261
Cohort 1C	8/11	0.727	0.557 (0.557)	0.654 (0.678)	0.664	0.554 (0.583)	0.554 (0.583)	0.655	0.519 (0.519)	0.572 (0.617)	0.533	0.615 (0.625)	0.625 (0.625)	0.673
Cohort 1D	3/5	0.6	0.477 (0.477)	0.548 (0.562)	0.548	0.444 (0.455)	0.444 (0.455)	0.542	0.448 (0.448)	0.516 (0.555)	0.480	0.512 (0.519)	0.519 (0.519)	0.543
			Basket 2 (Trastuzumab)											
Cohort 2A	9/24	0.375	0.365 (0.365)	0.395 (0.389)	0.379	0.371 (0.376)	0.371 (0.376)	0.386	0.367 (0.367)	0.381 (0.361)	0.362	0.367 (0.367)	0.367 (0.367)	0.371
Cohort 2B	11/24	0.458	0.401 (0.401)	0.463 (0.463)	0.430	0.392 (0.401)	0.392 (0.401)	0.422	0.406 (0.406)	0.396 (0.374)	0.374	0.406 (0.409)	0.409 (0.409)	0.425
Cohort 2C	4/19	0.211	0.278 (0.278)	0.274 (0.251)	0.277	0.291 (0.27)	0.291 (0.27)	0.315	0.295 (0.295)	0.354 (0.340)	0.340	0.295 (0.293)	0.293 (0.293)	0.264
Cohort 2D	3/8	0.375	0.354 (0.354)	0.420 (0.409)	0.381	0.357 (0.36)	0.357 (0.36)	0.382	0.365 (0.365)	0.383 (0.398)	0.361	0.365 (0.366)	0.366 (0.366)	0.371
			Basket 3 (Olaparib)											
Cohort 3A	14/23	0.609	0.508 (0.508)	0.587 (0.596)	0.568	0.517 (0.544)	0.517 (0.544)	0.6	0.475 (0.475)	0.496 (0.522)	0.473	0.536 (0.541)	0.541 (0.541)	0.570
Cohort 3B	10/24	0.417	0.405 (0.405)	0.429 (0.426)	0.413	0.385 (0.384)	0.385 (0.384)	0.382	0.388 (0.388)	0.415 (0.425)	0.401	0.405 (0.406)	0.406 (0.406)	0.407
Cohort 3C	8/25	0.32	0.334 (0.334)	0.350 (0.339)	0.337	0.336 (0.333)	0.336 (0.333)	0.339	0.344 (0.344)	0.374 (0.376)	0.364	0.343 (0.341)	0.341 (0.341)	0.336
Cohort 3D	3/17	0.177	0.252 (0.252)	0.253 (0.225)	0.248	0.261 (0.237)	0.261 (0.237)	0.269	0.288 (0.288)	0.326 (0.318)	0.320	0.261 (0.256)	0.256 (0.256)	0.225

Table 5.6: DRUP basket trials and the respective Cohorts masked. Under each utility function, the estimates are the scenario set 1 optimal mean and max aRMSE choice and in the parenthesis is the estimate based on the scenario set reduced optimal in mean and max aRMSE.

5.5 Discussion

This research study aimed to explore an approach to determine prior distribution and tuning parameter configurations essential for Bayesian methods used in estimating response rates within basket trial designs. The focus was on estimation, by developing a prior and tuning parameters selection method designed to be predefined and reproducible (table 5.4). Employing Simon's two-stage design in each independent cohort, we demonstrated our methodological framework using both simulation scenarios and real-world data from the DRUP study. Utility functions, specifically min-mean and min-max average Root Mean Square Error (aRMSE), determined the selection of optimal parameters performance. We applied the methods in a simulation environment where the study enrolled as planned 8 or 24 patient, as well as in a realistic to the DRUP study setting, where a reduced number of patients are observed in the second stage for some cohorts at a point when estimation becomes relevant.

A key strength of our approach is the systematic and transparent methodology for parameter optimization, allowing pre-specification and reproducibility of optimal choices based on clear performance metrics. Prior and tuning parameter selection is often arbitrary or unclear in existing Bayesian basket trial literature. Explicit distinctions between 'prior parameters' and 'tuning parameters' should be clarified, given their substantial influence on model behaviour. Additionally, by applying this approach directly to realistic scenarios from the DRUP study, we demonstrated its practical relevance and potential utility in real-world oncology trials.

Overall, we found that optimal parameter selections based on mean aRMSE and max aRMSE generally converged, producing similar estimated response rates across methods in the DRUP application (table 5.6). Although parameter patterns showed similarities across the utility functions, small differences were still influential for specific estimation outcomes. This emphasizes the importance of explicitly defining and justifying the utility criterion at the design stage.

Each method displayed distinct characteristics. The Berry method, provided generally low values for mean and maximum aRMSE, indicating good average performance. However, in heterogeneous conditions, it may not be the best choice. The EXNEX, although slightly under-performing in homogeneous scenarios (e.g., scenarios a, k, and o), consistently provided reliable and robust estimates in scenarios with higher heterogeneity. Psioda demonstrated parameter regions with low aRMSE, but with more sensitivity to the scenarios.The Fujikawa method showed consistent performance and stability in parameter selection across different scenario sets, making it a reliable choice for a range of trial conditions. The sample proportion is known to be biased for these STS designs, but is nevertheless often used. Our results indicate that borrowing information across baskets generally reduces estimation errors. Jung et. al introduced an UMVUE to provide an unbiased estimator for the STS design, as an alternative to the biased sample proportion. However, it assumes that each cohort reaches the full sample size which is not always the case in practice. Simulations in Scenario set 1 using the Jung estimator show that the mean aRMSE is 0.1 and the max aRMSE is 0.107.

The pooled estimate, commonly used in clinical practice, calculates the average of observed patient responses divided by the total number of available patients across all cohorts in the same basket. We applied the pooled estimate to our simulation study (Scenario Set 1), resulting in a mean aRMSE of 0.14, a maximum aRMSE of 0.26, and an average bias (aBias) of 0.12. However, we caution that this estimate's assumption, that the populations across cohorts are homogeneous, complicates the interpretation of bias and limits direct comparability with methods that explicitly model heterogeneity.

The sensitivity analysis underscored method-specific considerations, emphasizing implications rather than direct comparisons. Scenario set 2, characterized by higher true response probabilities (0.4, 0.7, 0.9), resulted in improved accuracy and reduced aRMSE. This improvement arises from the trial design's null hypothesis of 0.1, as

higher true response rates reduce early termination in the first stage, thereby reaching full planned enrollment and enhancing estimator precision. The Psioda, Fujikawa and Berry methods, were sensitive to prior mean choices, whereas EXNEX displayed relative robustness within its limited parameter exploration. Thus, choosing a Bayesian method necessitates careful consideration of sensitivity to the prior and realistic trial scenarios during the design phase.

The process of determining optimal parameters involves inherent limitations. Firstly, despite small observed differences in parameter space, selecting a single optimal setting remains challenging. This challenge can be overcome by sufficiently large simulation studies to reduce the impact of inter-simulation variability, thereby enabling robust and reliable parameter selection. Additionally, our use of min-mean and min-max aRMSE utility functions, although logically justified and operationally sound, inherently involves trade-offs. Averaging performance across diverse scenarios may mask extreme estimation errors specific to certain scenarios. Conversely, focusing on maximum errors might emphasize outliers, potentially sacrificing generalizability. Alternative or combined utility metrics could yield different optimal parameters, and thus careful consideration and explicit justification of the chosen metrics are crucial. Another consideration is the scenario set design itself. If future researchers adopt similar approaches, careful scenario and parameter-space definition at the design stage is essential. Realistic scenario representation, balanced across plausible trial outcomes, impacts both optimal parameter selection and the robustness of estimation. Hence, scenario choice itself is a critical design decision and a potential limitation if inadequately addressed. The computational burden of our proposed method restricted the number of Bayesian methods explored, compared to our previous work [70].

The simulation study provides valuable insights into the performance of different methods for basket trial analysis under the DRUP design. The approach described allows for rational prior and tuning parameter choice at the design stage of a basket trial. Future research should aim at refining these optimization methodologies to enhance robustness across wider scenario ranges and utility functions. Additional information from other even more complex endpoints could also be included in a more robust development of a basket trial estimator. Developing clear, practical guidelines for prior distribution and parameter selection in Bayesian basket trial models can substantially increase their applicability and reliability in precision medicine.

Appendix A

	Basket A	Basket B	Basket C	Basket D
a	0.9	0.9	0.9	0.9
b	0.9	0.9	0.9	0.7
c	0.9	0.9	0.9	0.4
d	0.9	0.9	0.7	0.7
e	0.9	0.9	0.7	0.4
f	0.9	0.9	0.4	0.4
g	0.9	0.7	0.7	0.7
h	0.9	0.7	0.7	0.4
i	0.9	0.7	0.4	0.4
j	0.9	0.4	0.4	0.4
k	0.7	0.7	0.7	0.7
l	0.7	0.7	0.7	0.4
m	0.7	0.7	0.4	0.4
n	0.7	0.4	0.4	0.4
o	0.4	0.4	0.4	0.4

Table 5.7: Scenarios set 2 which is used in the validation and comparison of the methods outcome

Appendix B

5.5.1 Scenario set 1

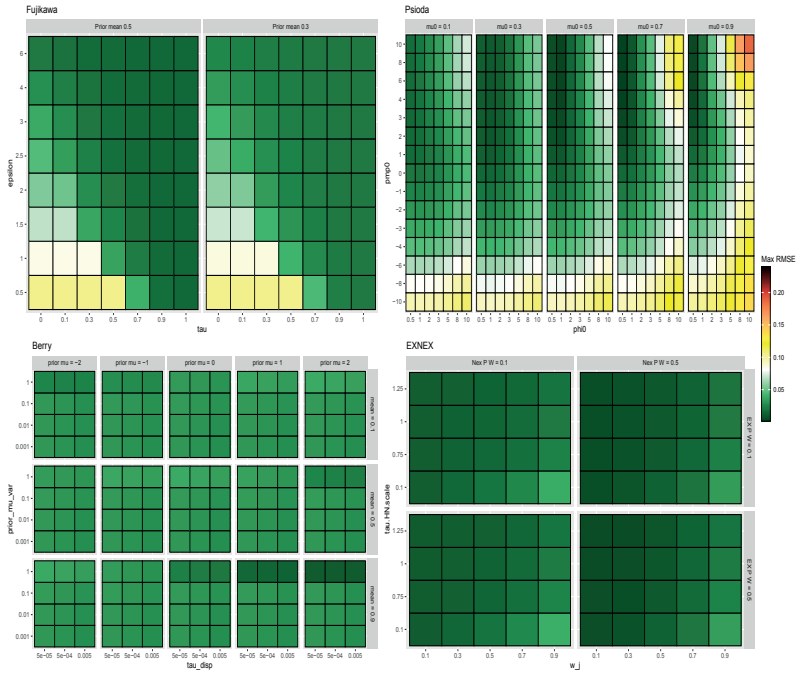


Figure 5.3: Results of the **scenario set 1**. A heat-map figure of the **aBias** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aBias values and colour combination.

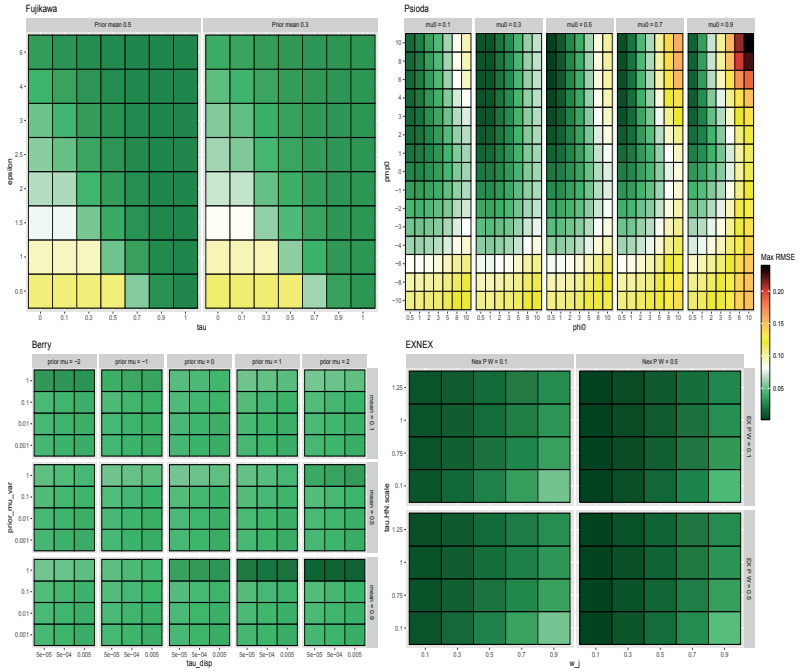


Figure 5.4: Results of the **scenario set reduced** number of patients. A heat-map figure of the **aBias** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aBias values and colour combination.

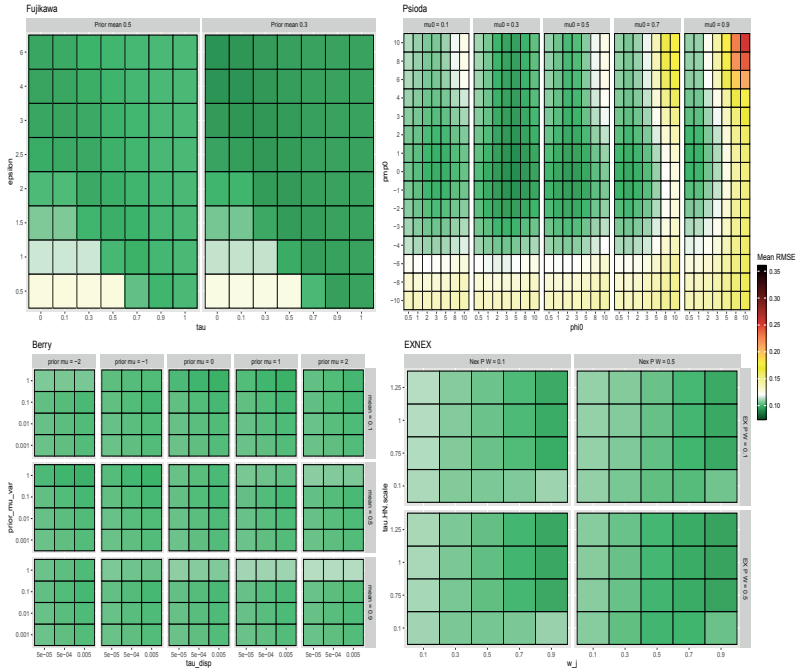


Figure 5.5: Results of the **scenario set reduced** number of patients. A heat-map figure of the **Mean aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

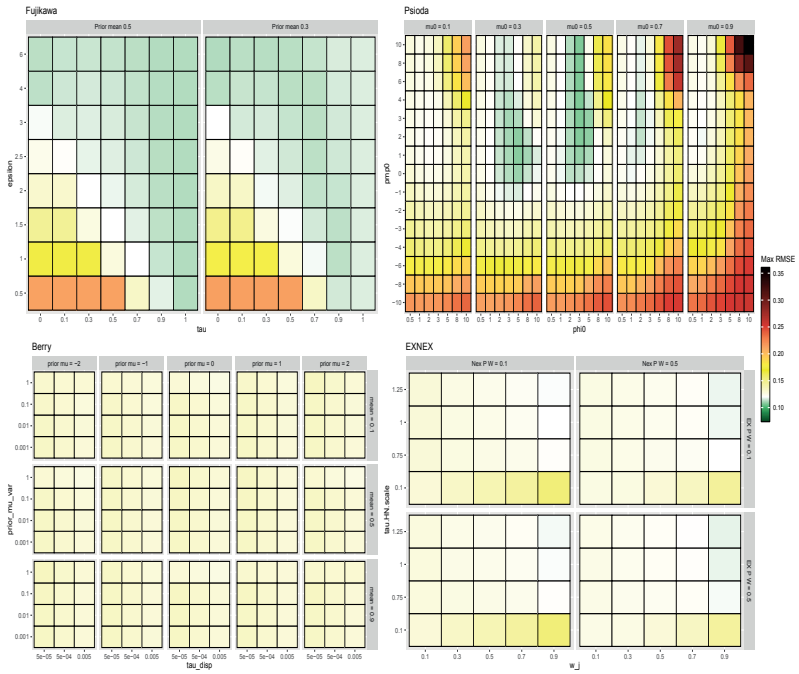


Figure 5.6: Results of the **scenario set reduced** number of patients. A heat-map figure of the **Max aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

5.5.2 Scenario set 2

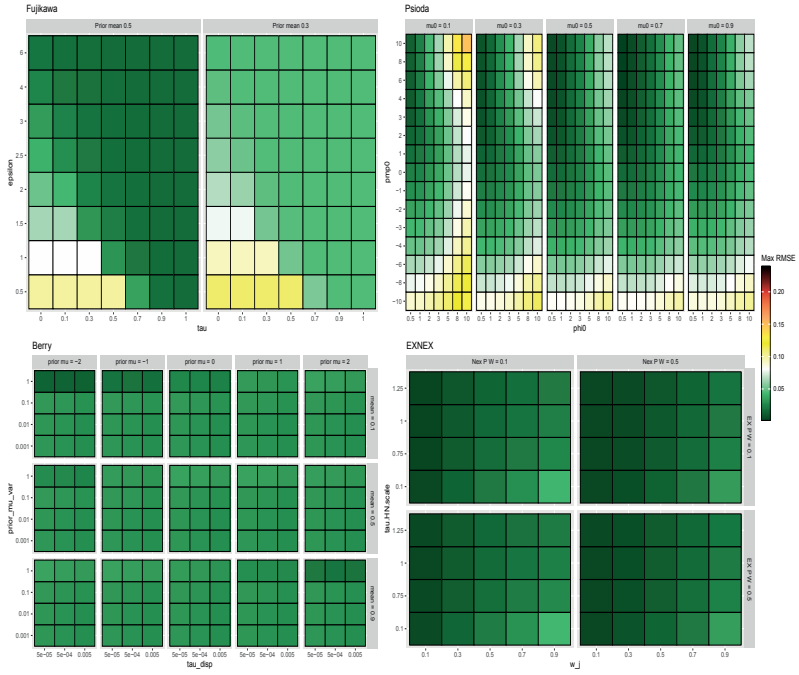


Figure 5.7: Results of the **scenario set 2**. A heat-map figure of the **aBias** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aBias values and colour combination.

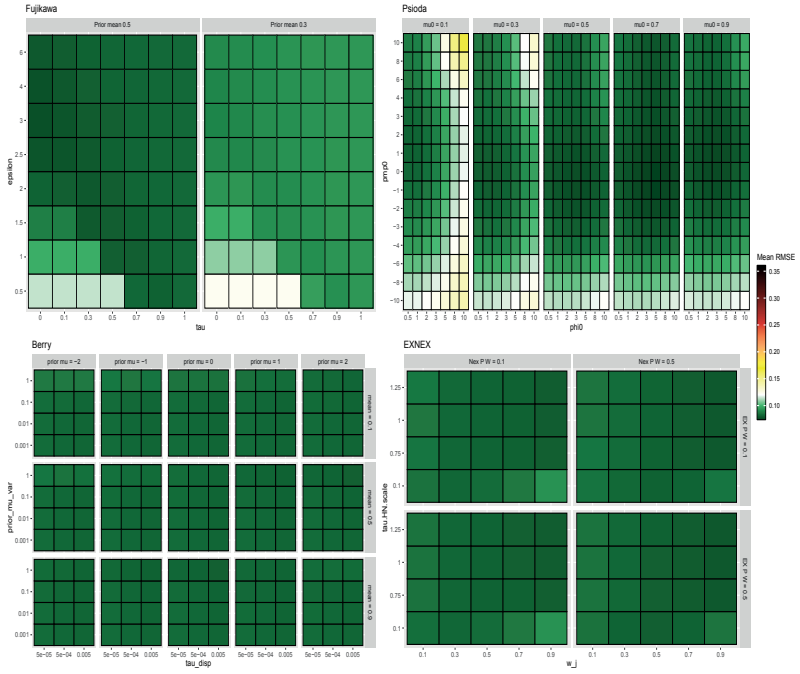


Figure 5.8: Results of the **scenario set 2**. A heat-map figure of the **Mean aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

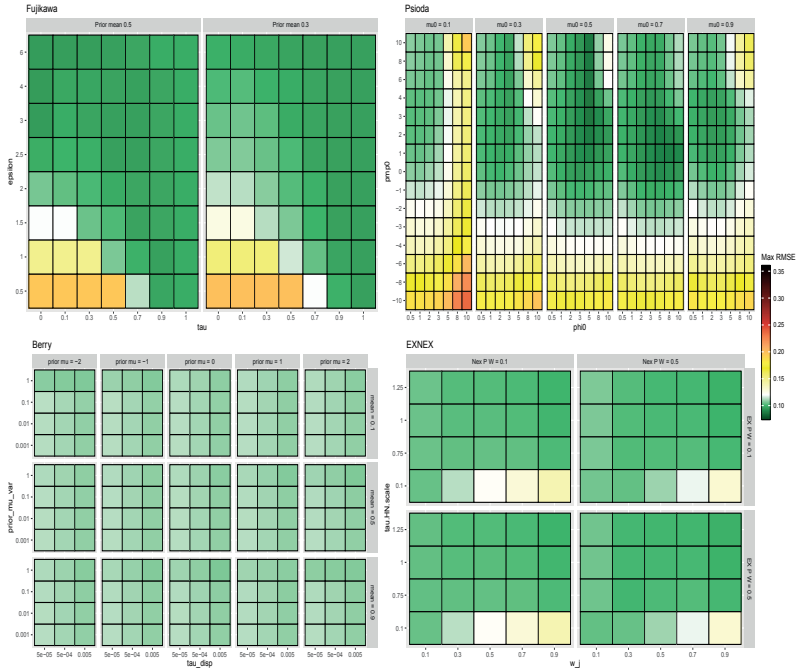


Figure 5.9: Results of the **scenario set 2**. A heat-map figure of the **Max aRMSE** for all the 4 methods, representing all the parameter combinations used. The bar on the side of the graph indicates the aRMSE values and colour combination.

Appendix C

Method	Parameter	$\hat{U}_{\text{min-mean aRMSE}}$	$\hat{U}_{\text{min-mean aRMSE}}$ set 2	$\hat{U}_{\text{min-max aRMSE}}$	$\hat{U}_{\text{min-max aRMSE}}$ set 2	Uninformative
Fujikawa	τ	0	0	0	0	0.5
	ϵ	4	3	4	4	2
	α	1	1	1	1	1
	β	2.333	1	2.333	1	1
Psioda	$\phi 0$	5	5	5	5	2
	$pmp0$	0	0	4	3	1
	$\mu 0$	0.3	0.7	0.5	0.5	0.5
Berry	$mean$	0.5	0.5	0.5	0.9	0.5
	$prior_{\mu.mean}$	-2	2	1	0	0
	$prior_{\mu.var}$	1	1	1	1	0.01
	τ_{disp}	0.005	0.005	0.005	0.005	0.0005
EXNEX	$ex.p.w$	0.5	0.5	0.5	0.5	0.5
	$nex.p.w$	0.5	0.5	0.5	0.5	0.5
	w	0.9	0.9	0.9	0.9	0.5
	$tau.HN.scale$	1	1	1.25	1.25	1

Table 5.8: Parameters optimal values on the different utility functions and the uninformative case, including both simulation set 1 and 2.

6

General discussion

Estimation of treatment effects in clinical trials connects the trial design with real-world applications beyond merely rejecting a null hypothesis of "no effect". This is even more relevant in the era of precision medicine and personalized treatments. Modern clinical trials, including basket trials and adaptive designs, are becoming increasingly complex, highlighting the need for tailored estimation methods. Good estimation methods are essential for evaluating the true effectiveness of new therapies and for guiding healthcare decisions that can affect patient outcomes and overall treatment strategies. With recent improvements in statistical techniques, we can now better understand the varied responses among different patient groups. This makes accurate estimation even more important, ensuring trial results are meaningful and applicable across diverse clinical trial designs.

Basket trials and master protocol designs represent an important evolution in clinical research methodology, particularly within oncology [10]. These trial frameworks are distinct in their ability to simultaneously explore multiple hypotheses, treatments, or patient groups within a single study, which significantly enhances research efficiency and the speed of drug development. Master protocols, including basket, umbrella, and platform trials, have gained prominence due to their potential to personalize medicine and optimize resource use, especially in complex diseases and rare genetic conditions.

However, the uptake of these innovative designs has not come without challenges. A primary difficulty identified across recent literature is the lack of consistent and standardized definitions and outcome reporting. This inconsistency complicates efforts to synthesize and compare trial results, creating barriers to broader implementation and regulatory acceptance, Siden et al. [73]. A notable gap remains between theoretical potential and practical implementation. Zhou and Ji [20] highlight that Bayesian statistical techniques can significantly enhance the precision and statistical power of basket trials by borrowing information across subgroups, yet their adoption remains limited. Kasim et al. [17] further point out that simpler, traditional trial designs still predominate, suggesting a hesitation or lack of readiness within the research community to fully embrace more complex but beneficial statistical methodologies. Operational complexities further challenge the effective deployment of master protocols. The management of multi-group patient enrollment, adaptive trial modifications, and appropriate use of historical or external control data demands substantial logistical and methodological expertise. Bofill et al. [74] underscore these challenges, specifically noting potential biases introduced when non-concurrent controls are improperly handled, which can adversely affect the validity of trial conclusions.

A notable complexity arises from Bayesian methods employed in basket trial analyses, such as hierarchical

models and adaptive borrowing techniques [15], [18], [54]–[59], [62], [64], [65], [72], [75]–[78]. Although these Bayesian approaches can be powerful, they introduce challenges, such as including computational intensity, parameter tuning complexities, and the subjective nature of prior distribution selection. In basket trials, choosing these parameters effectively implies assumptions about the similarity or heterogeneity of treatment effects across patient groups, and the anticipated treatment response characteristics within each basket. Thus, careful, pragmatic and biologically informed decisions about these parameters before trial initiation are crucial. However, current literature often lacks comprehensive guidance on optimal parameter selection, highlighting an important area for further methodological clarity and practical improvement in basket trial designs. While a practical distinction is often made between "prior parameters" and "tuning parameters," this separation is arguably artificial: both are pre-specified and influence model behavior in ways that reflect assumptions about the data. Should these two categories be unified under a broader conception of prior assumptions, and how might that shift affect the transparency and interpretability of Bayesian basket trial designs? Against this background, estimation in complex trials in oncology emerged as a relevant research question. Furthermore, the DRUP study [43] presents an important example of such a complex design, including multiple baskets and treatments and platform elements. This ongoing study gave rise to these questions, and was thus used as application for the methodology. In light of the basket trial design, estimation methods were addressed at the level of individual cohorts (substudies), as well as at the basket level.

In this thesis I systematically addressed statistical and methodological challenges associated with estimation in innovative clinical trial designs, with a particular focus on oncology. Across the four chapters, several insights and methodological contributions emerged, each addressing central issues in today's evolving clinical research landscape. These include improving estimation procedures under early stopping in two-stage designs, developing robust methods for censored survival data, and evaluating Bayesian approaches for information borrowing in basket trials. Collectively, this work contributes to a deeper understanding of how estimation interacts with trial design, and offers practical strategies for achieving more reliable and interpretable results in complex trial settings.

In **Chapter 2**, I address the need for improved estimation methods in STS trial designs when early stopping is applied. I focus on a specific design variant, the stopped STS (SSTS), where the trial stops as soon as a decision for efficacy or futility can be made, potentially before reaching the planned sample size. I develop a UMVUE for the response rate in the SSTS design and show that it maintains unbiasedness despite the flexible stopping rules. I also propose a method to construct exact confidence intervals based on sample space orderings, which aligns with the sequential nature of the design. Through simulation studies, I compare the SSTS UMVUE with standard estimators from the STS framework, evaluating their performance in terms of bias, mean squared error, coverage, and expected sample size. I demonstrate that while early stopping can reduce precision, it can also lead in time and resources efficiency and patient benefit.

In **Chapter 3**, I investigate the estimation of the Restricted Mean Duration of Response (RMDoR) in oncology. I focus on RMDoR as an alternative to the expected Duration of Response (DoR), particularly when right-censoring makes the latter difficult to estimate. Response durations are typically determined based on imaging of tumors at scheduled intervals, and because assessment schedules can differ between studies, comparing results across studies becomes challenging due to varying degrees of interval censoring. Additionally, the length of patient follow-up may vary substantially between studies, further complicating comparisons and potentially introducing bias. I examine how RMDoR behaves as a function of the truncation time τ, and explore how its interpretation depends on the choice of τ. I define multiple estimators for the RMDoR, considering practical aspects of interval censoring, either by using detection times directly or correcting with midpoint imputation. Through extensive simulation

studies, I assess the performance of these estimators in both single-arm and two-arm randomized trials under various scanning schedules and realistic survival scenarios. My results show that while interval censoring introduces bias in survival estimates, this has only a minor effect on RMDoR estimation. I highlight that the choice of τ plays a much larger role, especially for comparisons across treatments. I advocate for careful reporting of RMDoR estimates as a function of τ, and propose that RMDoR ratios may offer a stable and interpretable measure of treatment efficacy.

The PASKWIL-criteria 2021 [28] assess the clinical relevance of non-randomized studies by evaluating the lower bound of the 95%-CI of the objective response rate (ORR), paired with a point estimate of the median DoR. This framework does not account for the uncertainty in the DoR estimation, nor does it reflect the total patient population when response is rare. We contribute to this discussion by exploring the RMDoR as an alternative efficacy measure that incorporates both responders and non-responders, consistent with the updated estimand perspective proposed by Huang et al. [50]. We show that RMDoR can be reliably estimated even under interval censoring and right-censoring, and investigate how its behavior varies with respect to the choice of τ. To use RMDoR in decision-making, we recommend setting clear and interpretable thresholds. These could be based on existing clinical guidelines (e.g. translating PASKWIL's median DoR cutoffs into RMDoR equivalents), historical data, or expert consensus on what constitutes meaningful benefit. Including RMDoR alongside ORR may provide a more complete picture of treatment efficacy, and help refine clinical guidelines such as PASKWIL by explicitly linking treatment benefit to both effect size and estimation uncertainty.

In **Chapter 4**, I examine the problem of estimating cohort-specific response rates in single-stage basket trials, where a single treatment is evaluated across diverse cancer types sharing a molecular target. These designs, while operationally efficient, introduce statistical challenges when it comes to inference, particularly under the Bayesian framework which allows for information borrowing across cohorts. The primary aim of this chapter is to assess the accuracy and robustness of various Bayesian estimators that implement such borrowing, compared to the standard frequentist sample proportion. Using a comprehensive simulation study, I compare seven Bayesian estimators across a wide array of scenarios varying in response rate heterogeneity and sample size. I focus on three performance criteria: average absolute bias, mean squared error (MSE), and the degree of shrinkage towards the overall mean response. I show that in settings with little to no heterogeneity, the estimator proposed by Berry et al. achieves the lowest bias and MSE, benefiting from strong borrowing. However, in more heterogeneous settings, no single estimator consistently outperforms others. Instead, performance becomes highly context-dependent, influenced by model priors, tuning parameters, and the nature of between-cohort variation. My findings highlight the trade-off inherent in borrowing: while it can reduce variance, it may also introduce bias, especially when true response rates differ substantially between cohorts. Importantly, I reveal how prior specification, particularly the prior mean, can shift posterior estimates and impact conclusions. This underscores the need for careful prior elicitation in practice. Ultimately, I advocate for a context-aware selection of estimation methods in basket trials, encouraging the use of simulations during trial planning to guide the choice of model and hyperparameters. The insights from this work aim to inform researchers designing early-phase basket trials, where accurate estimation of response rates is crucial for deciding which combinations of drug and tumor types warrant further investigation.

In **Chapter 5**, I provide a simulations based approach to guide the choice of model and tuning parameters that can be used at the planning stage. To ensure applicability to the details of, e.g., the DRUP study design, I extended the evaluation of response rate estimation to basket trials that follow Simon's two-stage (STS) design, using selected examples from the DRUP study. Compared to single-stage settings, the two-stage design introduces additional complexities, particularly when patient accrual is incomplete or early stopping occurs. This chapter focuses on four Bayesian estimation approaches, Berry, EXNEX, Psioda, and Fujikawa, that facilitate information

sharing across cohorts. A key contribution of this work is the use of simulation-based parameter tuning, that can be used at the design stage. By exploring a broad grid of hyperparameter values, I identify optimal configurations under two utility functions: the minimization of the mean absolute Root Mean Square Error (min-mean aRMSE) and the minimization of the worst-case scenario error (min-max aRMSE). These choices reflect different priorities: achieving accurate estimation on average versus avoiding extreme errors. I also consider a reduced sample size scenario, simulating situations where the second stage of a cohort is not fully accrued. The results show that, in terms of these utility functions, all four Bayesian methods can outperform the frequentist sample proportion and UMVUE estimators, particularly under incomplete accrual. All methods could provide parameter sets where the aRMSE performance was strong. Berry, Fujikawa and Psioda, proved sensitive to the choice of priors in some scenarios, highlighting the importance of careful parameter tuning. Overall, EXNEX was most stable across all parameter spaces and scenario sets, though it should be noted that the parameter space explored for EXNEX was not as extensive as for the other methods. By applying these methods to actual DRUP data, I demonstrate how different estimators yield varying results across cohorts. This has direct implications for clinical interpretation. For example, in cohorts with high observed response, borrowing can shrink estimates downward, conversely, in smaller or uncertain cohorts, borrowing can lead to more optimistic estimates. These dynamics highlight the need for pre-specifying estimation strategies in basket trial protocols. This chapter shows that Bayesian borrowing is not a universal solution, but a flexible strategy that needs careful tuning of parameters. While simulations take time and computing power, they help identify the best settings for accurate estimates in multi-cohort trials with two-stage designs. These findings aim to help researchers working on basket trials especially in precision oncology make more informed choices at the design stage by balancing accuracy, consistency, and clarity.

One of the main conclusions of this thesis is that estimation and trial design are deeply connected in the context of basket trials. The type of design, whether single-stage, Simon's two-stage, or an adaptive structure, not only frames the trial, but also directly impacts how estimation should be performed. Estimation is not just something that follows data collection, it is embedded within the design itself, shaped by features like sample size restrictions, early stopping criteria, and the expected variability between cohorts.

This becomes particularly clear when estimation takes place under incomplete accrual, which is often the case in trials targeting rare patient groups. In such situations, standard estimators like the sample proportion or even the UMVUE may fail to perform adequately. As shown in this work, Bayesian estimators that borrow information across cohorts can provide more stable results, though they depend strongly on prior and tuning parameter choices. These parameters are not neutral they reflect assumptions about how similar or different the response rates are expected to be across baskets.

This introduces a broader methodological challenge: how to define and justify these assumptions clearly and transparently, ideally before the trial begins. For example, in the simulation studies, EXNEX appeared more robust across a wide range of scenarios, but this may be due to the limited parameter space, which could make the method appear more reliable than it actually is. On the other hand, methods like Psioda, Fujikawa, and Berry allowed for greater flexibility, but also showed more sensitivity to parameter tuning highlighting the trade-off between robustness and adaptability.

This suggests that parameter tuning should be considered part of the trial design process itself, rather than something addressed only during analysis. Although simulations can be computationally intensive, they enable researchers to identify optimal configurations tailored to the specific goals and constraints of a trial. Just as we predefine sample sizes or stopping rules, estimation strategies and tuning procedures should be incorporated into the protocol from the start.

At the same time, this work underlines a lack of formal guidance around the use of Bayesian estimators in trial designs. Regulatory documents like the ICH E9(R1) focus on the estimand framework and handling of intercurrent events, but give less attention to how Bayesian methods, particularly those sensitive to prior and tuning choices, fit within this framework. The distinction often made between "prior parameters" and "tuning parameters" may not be meaningful in practice, since both influence the final estimates. Treating them jointly as part of the prior structure could improve transparency and facilitate clearer reporting and justification.

Overall, this work points to a broader tension between simplicity and realism. Simple estimators are easy to apply, understand, and justify, but may overlook key sources of uncertainty or variation. Bayesian approaches, when well-calibrated, can offer a compromise: improving stability in estimation for small cohorts while still accounting for heterogeneity. However, their value depends on transparent assumptions, thoughtful design, and a clear understanding of their limitations.

This thesis has several limitations that should be acknowledged. The study design in basket trials was restricted to estimation under single-stage and Simon's two-stage designs, without exploring alternative approaches such as more complex Bayesian or adaptive trial designs. The range of estimation methods considered was limited by computational feasibility, as many advanced methods require significant resources and time, which also constrained the breadth of the parameter space examined in Chapter 5 and limited the generalizability of the results. Additionally, only one Simon's two-stage design configuration was investigated, and no comparison was made regarding the efficacy of estimators across different design choices. The scenarios and the number of cohorts used in the simulation studies were limited due to computational constraints, which prevented assessment of how the proposed methods perform under greater heterogeneity or with larger numbers of cohorts, an important consideration in practice. The procedure for optimizing Bayesian parameters followed a single specific approach and did not include comparisons with alternative parameter search strategies, potentially limiting recommendations for parameter tuning. Furthermore, the simulations in chapter 5 were primarily based on complete cohorts, whereas in real clinical trials, recruitment is often incomplete, especially in rare disease contexts, which may require analyses that deviate from the original design. Taken together, these limitations highlight the need for further research on a wider variety of designs, methods, and real-world data complexities to enhance the practical applicability and robustness of statistical approaches in basket trials.

Future research directions

This work opens several paths for future research that could further improve the design and analysis of basket trials. A central area of development lies in the use of more complex endpoints beyond binary outcomes. While response rate remains a commonly used measure in early-phase oncology trials, it inevitably discards part of the available information, especially regarding the timing and durability of response. Continuous endpoints, or combination of response and Time-to-event endpoints such as progression-free survival (PFS), duration of response (DoR), or even the restricted mean survival time (RMST) could offer richer insights if adapted appropriately to the basket trial context. Exploring robust estimation techniques for these endpoints, particularly under censoring and small sample constraints, could allow for more nuanced decision-making.

Another direction is the development of trial designs that integrate both treatment activity thresholds and estimation performance as core elements. Many current designs focus heavily on decision rules (e.g., Simon's two-stage) while treating estimation as secondary. However, as this thesis shows, estimation is part of the design, and separating the two can lead to inefficiencies or misleading conclusions. Future work could explore designs that optimize both decision-making and estimation accuracy, while still maintaining interpretability and feasibility in

rare populations. In addition, more clarity is needed around the role of Bayesian prior parameters. Their influence extends beyond just estimation, they shape the trial's behavior and outcomes from the design phase. Understanding this dual role and investigating whether common prior structures could serve both design and estimation objectives would be a valuable contribution to the methodology of basket trials.

Methodological advances in estimation procedures also remain a critical need. While this thesis evaluated several Bayesian estimators and highlighted the importance of prior and tuning parameter selection, there is room for improvement. New methods that are less sensitive to prior mis-specification, more computationally efficient, or better suited to small-sample inference would be highly valuable. Likewise, adapting techniques from machine learning or shrinkage-based approaches may help in producing more stable estimates while maintaining transparency.

Finally, there is a strong need for clearer guidance on how basket trials should be designed, analyzed, and reported when using advanced estimation methods. At present, many methodological decisions are left implicit or are made ad hoc. Developing best practice recommendations covering issues such as prior specification, tuning parameter grids, choice of estimators, and handling of incomplete data would help researchers plan more robust trials. In particular, trial protocols should clearly specify the estimation strategy alongside other key design elements, ensuring alignment with the estimand framework and regulatory expectations.

By addressing these areas, future work can build on the findings of this thesis to make basket trials more efficient, reliable, and impactful in precision oncology and beyond.

Bibliography

[1] R. Collier and Legumes, "Lemons and streptomycin: A short history of the clinical trial.," *CMAJ*, vol. 180, pp. 23–24, 2009. DOI: 10.1503/cmaj.081879.

[2] S. J. Dodgson, "The evolution of clinical trials.," *The Journal of the European Medical Writers Association*, vol. 15, pp. 20–21, 2006. [Online]. Available: https://www.researchgate.net/profile/Susanna-Dodgson-2/publication/242561119_The_evolution_of_clinical_trials/links/56a25f7c08ae232fb2019e62/The-evolution-of-clinical-trials.pdf.

[3] MRC Streptomycin in Tuberculosis Trials Committee, "Streptomycin treatment of pulmonary tuberculosis.," *BMJ*, vol. 2, pp. 769–783, 1948.

[4] S. E. Jackson and J. D. Chester, "Personalised cancer medicine.," *Int J Cancer*, vol. 137, no. 2, pp. 262–266, 2015. DOI: 10.1002/ijc.28940. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ijc.28940.

[5] E. A. Ashley, "Towards precision medicine.," *Nat Rev Genet*, vol. 17, pp. 507–522, 2016. DOI: 10.1038/nrg.2016.86. [Online]. Available: https://doi.org/10.1038/nrg.2016.86.

[6] M. Akacha, F. Bretz, and S. Ruberg, "Estimands in clinical trials broadening the perspective.," *Statist. Med.*, vol. 36, pp. 5–19, 2017. DOI: 10.1002/sim.7033. [Online]. Available: https://doi.org/10.1002/sim.7033.

[7] O. Collignon, C. Gartner, A. B. Haidich, *et al.*, "Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials.," *Clin. Pharmacol. Ther.*, vol. 107, pp. 1059–1067, 2020. DOI: 10.1002/cpt.1804. [Online]. Available: https://doi.org/10.1002/cpt.1804.

[8] I. C. F. H. of Technical Requirements For Pharmaceuticals For Human Use (ICH), "Ich e9(r1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials," 2019. [Online]. Available: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.

[9] H.-G. Eichler and F. Sweeney, "The evolution of clinical trials: Can we address the challenges of the future?" *Clinical Trials*, vol. 15, pp. 27–32, 2018. DOI: 10.1177/1740774518755058. [Online]. Available: https://doi.org/10.1177/1740774518755058.

[10] J. Woodcock and L. LaVange, "Master protocols to study multiple therapies, multiple diseases, or both," *New England Journal of Medicine*, vol. 377, pp. 62–70, 2017. DOI: 10.1056/NEJMra1510062. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra1510062.

[11] J. J. H. Park, E. Siden, M. J. Zoratti, *et al.*, "Systematic review of basket trials, umbrella trials, and platform trials: A landscape analysis of master protocols," *Trials*, vol. 20, p. 572, 2019. DOI: 10.1186/s13063-019-3664-1. [Online]. Available: https://doi.org/10.1186/s13063-019-3664-1.

[12] E. L. Meyer, P. Mesenbrink, C. Dunger-Baldauf, *et al.*, "The evolution of master protocol clinical trial designs: A systematic literature review," *Clinical Therapeutics*, vol. 42, pp. 1330–1360, 2020. DOI: 10.1016/j.clinthera.2020.05.010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32622783/.

[13] O. Collignon, A. Schiel, C. F. Burman, K. Rufibach, M. Posch, and F. Bretz, "Estimands and complex innovative designs.," *Clin. Pharmacol. Ther.*, vol. 112, pp. 1183–1190, 2022. DOI: 10.1002/cpt.2575. [Online]. Available: https://doi.org/10.1002/cpt.2575.

[14] M. Pohl, J. Krisam, and M. Kieser, "Categories, components, and techniques in a modular construction of basket trials for application and further research," *Biometrical Journal*, vol. 63, no. 6, pp. 1159–1184, 2021. DOI: 10.1002/bimj.202000314. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33942894/.

[15] X. Chi, Y. Yuan, Z. Yu, and R. Lin, "A generalized calibrated bayesian hierarchical modeling approach to basket trials with multiple endpoints," *Biom. J.*, vol. 66, p. 2, 2024. DOI: 10.1002/bimj.202300122. [Online]. Available: https://doi.org/10.1002/bimj.202300122.

[16] R. Simon, "Optimal two-stage designs for phase ii clinical trials," *Control Clinical Trials*, vol. 10, no. 1, pp. 1–10, 1989. DOI: 10.1016/0197-2456(89)90015-9. [Online]. Available: https://doi.org/10.1016/0197-2456(89)90015-9.

[17] A. Kasim, S. Bean, C. Hendriksen, H. Chen, Y. Zhou, and M. Psioda, "Basket trials in oncology: A systematic review of practices and methods, comparative analysis of innovative methods, and an appraisal of a missed opportunity," *Front. Oncol.*, 2023. DOI: 10.3389/fonc.2023.1266286. [Online]. Available: https://doi.org/10.3389/fonc.2023.1266286.

[18] L. Kanapka and A. Ivanova, "A frequentist design for basket trials using adaptive lasso," *Stat. Med.*, vol. 43, no. 1, pp. 156–172, 2024. DOI: 10.1002/sim.9947. [Online]. Available: https://doi.org/10.1002/sim.9947.

[19] K. Cunanan, A. Iasonos, R. Shen, C. B. Begg, and M. Gonen, "An efficient basket trial design," *Stat. Med.*, vol. 30, no. 10, pp. 1568–1579, 2011. DOI: 10.1002/sim.7227. [Online]. Available: https://doi.org/10.1002/sim.7227.

[20] T. Zhou and Y. Ji, "Bayesian methods for information borrowing in basket trials: An overview," *Cancers*, vol. 16, p. 251, 2024. DOI: 10.3390/cancers16020251. [Online]. Available: https://doi.org/10.3390/cancers16020251.

[21] J. Whitehead, "On the bias of maximum likelihood estimation following a sequential test," *Biometrika*, vol. 73, no. 3, pp. 573–581, 1986. DOI: 10.2307/2336521. [Online]. Available: https://doi.org/10.2307/2336521.

[22] S. H. Jung and K. M. Kim, "On the estimation of the binomial probability in multistage clinical trials," *Statistics in Medicine*, vol. 23, no. 6, pp. 881–896, 2004.

[23] A. O. Ayanlowo and D. T. Redden, "A two-stage conditional power adaptive design adjusting for treatment by covariate interaction," *Contemporary Clinical Trials*, vol. 29, pp. 428–438, 2008. DOI: 10.1016/j.cct.2007.10.003. [Online]. Available: https://doi.org/10.1016/j.cct.2007.10.003.

[24] C.-M. Chen and Y. Chi, "Curtailed two-stage designs with two dependent binary endpoints," *Pharmaceutical Statistics*, vol. 11, pp. 57–62, 2012. DOI: 10.1002/pst.496. [Online]. Available: https://doi.org/10.1002/pst.496.

[25] C. U. Kunz and M. Kieser, "Curtailment in single-arm two-stage phase ii oncology trials," *Biometrical Journal*, vol. 54, pp. 445–456, 2012. DOI: 10.1002/bimj.201100128. [Online]. Available: https://doi.org/10.1002/bimj.201100128.

[26] C. Hu, M. Wang, C. Wu, H. Zhou, C. Chen, and S. Diede, "Comparison of duration of response vs conventional response rates and progression free survival as efficacy end points in simulated immuno-oncology clinical trials," *JAMA Netw Open*, vol. 4, no. 5, e218175, 2021.

[27] H. J. Weber, S. Corson, J. Li, *et al.*, "Duration of and time to response in oncology clinical trials from the perspective of the estimand framework.," *Pharm. Stat.*, vol. 23, 2024. DOI: 10.1002/pst.2340. [Online]. Available: https://doi.org/10.1002/pst.2340.

[28] N. C. for the Assessment of Oncological Drugs (BOM), "Paskwil criteria," 2021. [Online]. Available: https://www.nvmo.org/over-de-adviezen/.

[29] T. Koyama and H. Chen, "Proper inference from simon's two-stage designs," *Statistics in Medicine*, vol. 27, no. 16, pp. 3145–3154, 2008. DOI: 10.1002/sim.3123. [Online]. Available: https://doi.org/10.1002/sim.3123.

[30] R. Porcher and K. Desseaux, "What inference for two-stage phase ii trials?" *BMC Medical Research Methodology*, vol. 12, p. 117, 2012. DOI: 10.1186/1471-2288-12-117. [Online]. Available: https://doi.org/10.1186/1471-2288-12-117.

[31] M. N. Chang, H. S. Wieand, and V. T. Chang, "The bias of the sample proportion following a group sequential phase ii clinical trials," *Statistics in Medicine*, vol. 8, no. 5, pp. 563–570, 1989.

[32] H. Y. Guo and A. Liu, "A simple and efficient bias-reduced estimator of response probability following a group sequential phase ii trial," *Journal of Biopharmaceutical Statistics*, vol. 15, no. 5, pp. 773–781, 2005. DOI: 10.1081/BIP-200067771. [Online]. Available: https://doi.org/10.1081/BIP-200067771.

[33] K. Kunzmann and M. Kieser, "Point estimation and p-values in phase ii adaptive two-stage designs with a binary endpoint," *Statistics in Medicine*, vol. 36, pp. 971–984, 2017. DOI: 10.1002/sim.7200. [Online]. Available: https://doi.org/10.1002/sim.7200.

[34] C.-M. Chen, "The uniformly minimum variance unbiased estimator of response rate in a two-stage design," *Journal of the Chinese Statistical Association*, vol. 53, pp. 262–273, 2015.

[35] Q. Li, "An mse-reduced estimator for the response proportion in a two-stage clinical trial," *Pharmaceutical Statistics*, vol. 10, pp. 277–279, 2011. DOI: 10.1002/pst.414. [Online]. Available: https://doi.org/10.1002/pst.414.

[36] M. S. Pepe, Z. Feng, G. Longton, and J. Koopmeiners, "Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility," *Statistics in Medicine*, vol. 28, no. 5, pp. 762–779, 2009. DOI: 10.1002/sim.3506. [Online]. Available: https://doi.org/10.1002/sim.3506.

[37] W.-Y. Tsai, Y. Chi, and C.-M. Chen, "Interval estimation of binomial proportion in clinical trials with a two-stage design," *Statistics in Medicine*, vol. 27, pp. 15–35, 2008. DOI: 10.1002/sim.2930. [Online]. Available: https://doi.org/10.1002/sim.2930.

[38] G. Shan, "Exact confidence limits for the probability of response in two-stage designs," *Statistics*, vol. 52, no. 5, pp. 1086–1095, 2018. DOI: 10.1080/02331888.2018.1469023. [Online]. Available: https://doi.org/10.1080/02331888.2018.1469023.

[39] C. Jennison and B. W. Turnbull, "Confidence intervals for a binomial parameter following a multi-stage test with application to mil-std 105d and medical trials," *Technometrics*, vol. 25, pp. 49–58, 1983.

[40] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404–413, 1934.

[41] J. Quispel-Janssen, V. van der Noort, J. F. de Vries, *et al.*, "Programmed death 1 blockade with nivolumab in patients with recurrent malignant pleural mesothelioma," *Journal of Thoracic Oncology*, vol. 13, no. 10, pp. 1569–1576, 2018. DOI: 10.1016/j.jtho.2018.05.038.

[42] P. K. Mangat, S. Halabi, S. S. Bruinooge, *et al.*, "Rationale and design of the targeted agent and profiling utilization registry study," *American Society of Clinical Oncology*, 2018. DOI: 10.1200/PO.18.00122. [Online]. Available: https://doi.org/10.1200/PO.18.00122.

[43] D. L. Van der Velden, L. R. Hoes, H. van der Wijngaart, *et al.*, "The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs," *Nature*, vol. 574, pp. 127–131, 2019. DOI: 10.1038/s41586-019-1600-x. [Online]. Available: https://doi.org/10.1038/s41586-019-1600-x.

[44] F. Bijma, M. A. Jonker, and A. W. van der Vaart, *An Introduction to Mathematical Statistics*. Amsterdam University Press, 2017.

[45] A. Delgado and A. Kumar, "Clinical endpoints in oncology - a primer," *Am J Cancer Res*, vol. 11, no. 4, pp. 1121–1131, 2021. [Online]. Available: www.ajcr.us%20/ISSN:2156-6976/ajcr0130927.

[46] T. Choueiri, R. Motzer, B. Rini, and et al., "Updated efficacy results from the javelin renal 101 trial: First-line avelumab plus axitinib versus sunitinib in patients with advanced renal cell carcinoma," *Ann Oncol*, vol. 31, no. 8, pp. 1030–1039, 2020.

[47] E. Saad, E. Coart, V. Deltuvaite-Thomas, L. Garcia-Barrado, T. Burzykowski, and M. Buyse, "Trial design for cancer immunotherapy: A methodological toolkit," *Cancers*, vol. 15, p. 4669, 2023.

[48] B. Huang, L. Tian, E. Talukder, M. Rothenberg, D. Kim, and L. Wei, "Evaluating treatment effect based on duration of response for a comparative oncology study," *JAMA Oncol*, vol. 4, no. 6, pp. 874–876, 2018.

[49] B. Huang, L. Tian, Z. McCaw, and et al., "Analysis of response data for assessing treatment effects in comparative clinical studies," *Ann Intern Med*, vol. 173, no. 5, pp. 368–374, 2020.

[50] B. Huang and L. Tian, "Utilizing restricted mean duration of response for efficacy evaluation of cancer treatments," *Pharmaceutical Statistics*, pp. 1–14, 2022.

[51] O. Aalen, O. Borgan, and H. Gjessing, *Survival and Event History Analysis*. Springer-Verlag, 2008.

[52] Z. Zhang, "Interval censoring," *SMMR*, vol. 19, pp. 53–70, 2010.

[53] European Medicines Agency, *Bavencio ema/chmp/550625/2019 epar assessment report*, Procedure No. EMEA/H/C/004338/II/0009/G, 2019.

[54] S. M. Berry, K. R. Broglio, S. Groshen, and D. A. Berry, "Bayesian hierarchical modelling of patient subpopulations: Efficient designs of phase in oncology clinical trials," *Clinical Trials*, vol. 10, pp. 720–734, 2013.

[55] B. Neuenschwander, S. Wandel, S. Roychoudhury, and S. Bailey, "Robust exchangeability designs for early phase clinical trials with multiple strata," *Pharmaceutical Statistics*, vol. 15, pp. 123–134, 2016. DOI: 10.1002/pst.1730. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/pst.1730.

[56] J. Jin, M.-K. Riviere, X. Luo, and Y. Dong, "Bayesian methods for the analysis of early-phase oncology basket trials with information borrowing across cancer types," *Statistics in Medicine*, vol. 39, pp. 3459–3475, 2020. DOI: 10.1002/sim.8675. [Online]. Available: https://doi.org/10.1002/sim.8675.

[57] M. A. Psioda, J. Xu, Q. Jiang, C. Ke, Z. Yang, and J. G. Ibrahim, "Bayesian adaptive basket trial design using model averaging," *Biostatistics*, vol. 22, pp. 19–34, 2019. DOI: 10.1093/biostatistics/kxz014. [Online]. Available: https://doi.org/10.1093/biostatistics/kxz014.

[58] K. Fujikawa, S. Teramukai, I. Yokota, and T. Daimon, "A bayesian basket trial design that borrows information across strata based on the similarity between the posterior distributions of the response probability," *Biometrical Journal*, vol. 62, pp. 330–338, 2020. DOI: 10.1002/bimj.201800404. [Online]. Available: https://doi.org/10.1002/bimj.201800404.

[59] N. Chen and J. J. Lee, "Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints," *Statistical Methods in Medical Research*, vol. 29, no. 9, pp. 2717–2732, 2020. DOI: 10.1177/0962280220910186. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32178585/.

[60] D. Blei, T. Griffiths, and M. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *ACM*, vol. 57, no. 2, 2010. DOI: 10.1145/1667053.1667056. [Online]. Available: https://cocosci.princeton.edu/tom/papers/ncrp.pdf.

[61] N. Chen and J. J. Lee, "Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes," *Biometrical Journal*, vol. 61, no. 5, pp. 1219–1231, 2019. DOI: 10.1002/bimj.201700275. [Online]. Available: https://doi.org/10.1002/bimj.201700275.

[62] Y. Liu, M. Kane, D. Esserman, D. Zelterman, and W. Wei, "Bayesian local exchangeability design for phase ii basket trials," *Statistics in Medicine*, 2022. DOI: 10.1002/sim.9514. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.9514.

[63] EMA, *Ema, assessment report*, n.d. [Online]. Available: https://www.ema.europa.eu/en/documents/assessment-report/keytruda-epar-public-assessment-report_en.pdf.

[64] R. Simon, S. Geyer, J. Subramanian, and S. Roychowdhury, "The bayesian basket design for genomic variant-driven phase ii trials," *Seminars in Oncology*, vol. 43, no. 1, pp. 13–18, 2016. DOI: 10.1053/j.seminoncol.2016.01.002. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26970120/.

[65] H. Zhou, F. Liu, C. Wu, E. H. Rubin, V. L. Giranda, and C. Chen, "Optimal two-stage designs for exploratory basket trials," *Contemporary Clinical Trials*, vol. 85, p. 105 807, 2019.

[66] X. Wu, C. Wu, F. Liu, H. Zhou, and C. Chen, "A generalized framework of optimal two-stage designs for exploratory basket trials," *Statistics in Biopharmaceutical Research*, 2021.

[67] Y. Jing, "An optimal two-stage exploratory basket trial design with aggregated futility analysis," *Contemporary Clinical Trials*, vol. 116, p. 106 741, 2022.

[68] S. H. Jung, T. Lee, K. Kim, and S. L. George, "Admissible two-stage designs for phase ii cancer clinical trials," *Statistics in Medicine*, vol. 23, no. 4, pp. 561–569, 2004. DOI: 10.1002/sim.1600.

[69] B. P. Hobbs, R. C. Pestana, E. C. Zabor, A. M. Kaizer, and D. S. Hong, "Basket trials: Review of current practice and innovations for future trials," *Journal of Clinical Oncology*, vol. 40, no. 30, pp. 3520–3528, 2022. DOI: 10.1200/JCO.21.02285. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10476732/.

[70] A. Daletzakis, R. van den Bor, V. van der Noort, and K. C. B. Roes, *Estimation in basket trials, evaluating the borrowing information performance, a methodological review*, 2024.

[71] *Utility-based optimization of fujikawa's basket trial design – pre-specified protocol of a comparison study*, Available online.

[72] *A basket trial design based on power priors*, Pre-print.

[73] E. G. Siden, J. J. Park, and M. J. e. a. Zoratti, "Reporting of master protocols towards a standardized approach: A systematic review," *Contemporary Clinical Trials Communications*, vol. 15, 2019. DOI: 10.1016/j.conctc.2019.100406. [Online]. Available: https://doi.org/10.1016/j.conctc.2019.100406.

[74] M. Bofill Roig, C. Burgwinkel, U. Garczarek, *et al.*, "On the use of non-concurrent controls in platform trials: A scoping review," *Trials*, vol. 24, 2023. DOI: 10.1186/s13063-023-07398-7. [Online]. Available: https://doi.org/10.1186/s13063-023-07398-7.

[75] K. Takeda, S. Liu, and A. Rong, "Constrained hierarchical bayesian model for latent subgroups in basket trials with two classifiers," *Statistics in Medicine*, vol. 41, no. 2, pp. 298–309, 2022. DOI: 10.1002/sim.9237. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/sim.9237.

[76] H. Zheng and J. M. S. Wason, "Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy," *Biostatistics*, vol. 23, no. 1, pp. 120–135, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8759447/.

[77] D. Kang, C. S. Coffey, B. J. Smith, Y. Yuan, Q. Shi, and J. Yin, "Hierarchical bayesian clustering design of multiple biomarker subgroups (hcombs)," *Statistics in Medicine*, vol. 40, no. 12, pp. 2893–2921, 2021. DOI: 10.1002/sim.8946. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33772843/.

[78] J. Asano, H. Sato, and A. Hirakawa, "Practical basket design for binary outcomes with control of family-wise error rate," *BMC Medical Research Methodology*, vol. 23, p. 52, 2023. DOI: 10.1186/s12874-023-01872-1. [Online]. Available: https://doi.org/10.1186/s12874-023-01872-1.

Samenvatting in het Nederlands

In dit proefschrift heb ik systematisch statistische en methodologische oplossing onderzocht voor het schatten van behandeleffecten in innovatieve klinische proefopzetten, met een bijzondere focus op oncologie. In de vier hoofdstukken komen verschillende inzichten en methodologische bijdragen naar voren, die specifiek problemen in het huidige, snel veranderende landschap van innovatieve designs van klinische studies adresseren. Deze omvatten het verbeteren van schattingsprocedures bij vroegtijdige stopzetting in het bekende Simon's Two-Stage (STS) design, het ontwikkelen van robuuste methoden voor het schatten van "duration of response" op basis van gecensureerde overlevingsdata, en het evalueren van Bayesiaanse benaderingen voor informatie-uitwisseling ('borrowing') in basket trials. Gezamenlijk draagt dit werk bij aan een dieper begrip van hoe schatting samenhangt met onderzoeksopzet, en biedt het praktische strategieën voor betrouwbaardere en beter interpreteerbare resultaten in complexe onderzoeksdesigns.

In **Hoofdstuk 2** behandel ik verbeterde schattingsmethoden in STS onderzoeksopzetten wanneer de studie voortijdig wordt gestopt. Ik richt mij op een specifieke variant, het gestopte STS (SSTS)-ontwerp, waarbij de studie wordt gestopt zodra een beslissing over werkzaamheid of futuliteit kan worden genomen, mogelijk nog vóór het bereiken van de geplande steekproefgrootte. Ik ontwikkel een uniform unbiased estimator (UMVUE) voor het responspercentage in dit SSTS-ontwerp en toon aan dat deze inderaad zuiver is, ondanks de flexibele stopregels. Ook stel ik een methode voor om exacte betrouwbaarheidsintervallen te construeren op basis van de steekproefruimte-ordening, wat aansluit bij het sequentiële karakter van het ontwerp. Via simulatiestudies vergelijk ik de SSTS UMVU met standaard schatters, op basis van bias, mean squared error, dekkingsgraad, en verwachte steekproefgrootte. Ik laat zien dat vroegtijdige stopzetting weliswaar precisie kan verminderen, maar ook kan leiden tot efficiënter gebruik van tijd en middelen, en tot voordeel voor patiënten.

In **Hoofdstuk 3** onderzoek ik de schatting van de Restricted Mean Duration of Response (RMDoR) in de oncologie. Ik focus op RMDoR als alternatief voor de verwachte duur van respons (DoR), met name wanneer rechtscensurering de DoR lastig te schatten maakt. Responsduur wordt doorgaans bepaald op basis van tumorbeeldvorming op geplande tijdstippen. Doordat beoordelingsschema's tussen studies kunnen verschillen, wordt vergelijking van resultaten bemoeilijkt omdat dit leidt tot door verschillen in intervalcensurenring. Ook kan de follow-up duur van patiënten sterk uiteenlopen tussen studies, wat vergelijkingen van resultaten tussen studies verder bemoeilijkt en mogelijk bias introduceert. Ik onderzoek hoe de RMDoR zich gedraagt als functie van de truncatietijd τ , en hoe de interpretatie afhangt van de keuze van τ . Ik definieer meerdere schatters voor de RMDoR, rekening houdend met praktische aspecten van intervalcensoring, hetzij door directe detectietijden te gebruiken, hetzij door correctie met mid-point imputatie. Met uitgebreide simulaties bestudeer ik de prestaties van deze schatters in zowel enkelarmige als gerandomiseerde studies onder diverse scanregimes en realistische overlevingsscenario's. Mijn resultaten tonen aan dat intervalcensoring weliswaar bias introduceert in overlevingsschattingen, maar dat het effect op RMDoR beperkt is. Ik onderstreep dat de keuze van τ een veel grotere rol speelt, vooral

voor vergelijkingen tussen behandelingen. Ik pleit voor zorgvuldige rapportage van RMDoR-schattingen als functie van τ , en stel voor dat RMDoR-ratio's een stabiele en interpreteerbare maat van behandelresultaat kunnen bieden. De PASKWIL-criteria 2021 [28] beoordelen de klinische relevantie van niet-gerandomiseerde studies aan de hand van de ondergrens van het 95%-betrouwbaarheidsinterval van het objectieve responspercentage (ORR), samen met een puntschatting van de mediane DoR. Dit raamwerk houdt geen rekening met de onzekerheid in de DoR-schatting, noch met de gehele populatie bij zeldzame respons. Ik draag bij aan deze discussie door de RMDoR te verkennen als alternatieve uitkomstmaat die zowel responders als non-responders omvat, in lijn met het aangepaste estimand-perspectief voorgesteld door Huang et al. [50]. Ik toon aan dat RMDoR betrouwbaar kan worden geschat, zelfs bij interval- en rechtscensoring, en laat zien hoe het gedrag varieert afhankelijk van τ . Voor gebruik van RMDoR in besluitvorming raden we aan om duidelijke en interpreteerbare drempelwaarden te formuleren, bijvoorbeeld door PASKWIL's DoR-drempels te vertalen naar RMDoR-equivalenten, historische data of expert consensus. Het toevoegen van RMDoR naast ORR kan een completer beeld geven van behandelingseffectiviteit en bijdragen aan verfijning van klinische richtlijnen zoals PASKWIL, door expliciet het voordeel te koppelen aan zowel effectgrootte als schattingsonzekerheid.

In **Hoofdstuk 4** onderzoek ik het schatten van cohort-specifieke responspercentages in basket trials, waarbij één behandeling wordt geëvalueerd in verschillende kankersoorten met een gedeeld moleculair target. Deze designs zijn operationeel efficient. Statistische efficiëntie winst is mogelijk, met name door Bayesiaanse methoden die informatie-uitwisseling tussen cohorten mogelijk maken. Het doel van dit hoofdstuk is het evalueren van de nauwkeurigheid en robuustheid van diverse Bayesiaanse schatters die deze 'borrowing' implementeren, vergeleken met de standaard frequentistische steekproef proportie. Met een uitgebreide simulatiestudie vergelijk ik zeven Bayesiaanse schatters in uiteenlopende scenario's van heterogeniteit in response percentage tussen cohorten en verschillende cohort steekproefgroottes. De evaluatie is gebaseerd op drie prestatiecriteria: gemiddelde absolute bias, mean squared error (MSE), en de mate van shrinkage richting het overall gemiddelde respons percentage. Ik laat zien dat bij geringe heterogeniteit de door Berry et al. voorgestelde schatter het kleinste bias en MSE bereikt dankzij sterke 'borrowing'. Echter, in secnario's met meer heterogeniteot tussen cohorten presteert geen enkele schatter consistent het best; de prestatie wordt context afhankelijk en beïnvloed door model priors, tuning parameters en de aard van variatie tussen cohorten. Mijn bevindingen illustreren de bekende "bias-variance trade-off" bij 'borrowing': het kan variantie verminderen, maar ook bias introduceren, vooral als de werkelijke respons percentages sterk verschillen tussen cohorten. Ik laat zien hoe specificatie van de priors, met name voor het gemiddelde, de posterior-schattingen en conclusies kan beïnvloeden. Dit benadrukt het belang van zorgvuldige prior-elicitatie in de praktijk. Ik bepleit een context afhankelijke keuze van schattingsmethoden in basket trials en het gebruik van simulaties bij de onderzoeksplanning om de keuze van model en. hyperparameters te sturen. De inzichten kunnen onderzoekers ondersteunen bij het ontwerpen van basket trials, waar nauwkeurige schatting van responspercentages essentieel is voor vervolgonderzoek naar veelbelovende behandelingen.

In **Hoofdstuk 5** volg ik een simulatie-gebaseerde aanpak om de keuze van model en tuning parameters te ondersteunen bij de onderzoeksplanning. Om toepasbaarheid te waarborgen voor de DRUP-studie, breidde ik de evaluatie van respons percentages uit naar basket trials die het Simon's two-stage (STS)-ontwerp volgen, aan de hand van geselecteerde voorbeelden uit de DRUP-studie. Dit hoofdstuk richt zich op vier Bayesiaanse schattingsmethoden (Berry, EXNEX, Psioda en Fujikawa) die informatie-uitwisseling tussen cohorten mogelijk maken. Een belangrijke bijdrage is het gebruik van simulatie-gebaseerde parameter keuze ('tuning'), bruikbaar tijdens de ontwerpfase. Door een breed raster van hyperparameter waarden te verkennen, identificeer ik optimale configuraties onder twee gedefinieerde nutsfuncties ("utilities"): het minimaliseren van de gemiddelde absolute Root Mean Squared Error (min-mean aRMSE) en het minimaliseren van de fout in het slechtst mogelijke scenario

(min-max aRMSE). Deze keuzes weerspiegelen verschillende prioriteiten: nauwkeurige schatting gemiddeld versus het vermijden van extreme fouten. Ook simuleer ik een scenario met door verkleinde steekproefgrootte, als de tweede fase van een cohort nog niet volledig kan worden geïncludeerd in de schatting van respons percantages. De resultaten laten zien dat, volgens deze nutsfuncties, alle vier de Bayesiaanse methoden beter kunnen presteren dan de frequentistische steekproefproportie en UMVUE, met name bij incomplete inclusie. Voor alle methoden konden parametersets worden gevonden met sterke aRMSE-prestaties. Berry, Fujikawa en Psioda bleken gevoelig voor de keuze van priors in sommige scenario's, wat het belang van zorgvuldige tuning onderstreept. Over het geheel genomen was EXNEX het meest stabiel over alle parameter- en scenario-ruimtes, hoewel het verkennings gebied voor EXNEX beperkter was dan voor de andere methoden. Door deze methoden toe te passen op daadwerkelijke DRUP-data, laat ik zien hoe verschillende schatters uiteenlopende resultaten geven per cohort. Dit heeft directe gevolgen voor de klinische interpretatie. In cohorten met een hoog waargenomen respons, kan 'borrowing' leiden tot lagere schattingen, terwijl in kleinere of onzekere cohorten het juist optimistisch kan uitpakken. Deze dynamiek benadrukt de noodzaak om de precieze methode van schatten vooraf te specificeren in het onderzoeksprotocol. Dit hoofdstuk toont aan dat Bayesiaans 'borrowing' geen universele oplossing is, maar een flexibele strategie die zorgvuldige keuzes van prior en tuning parameters vereist. Hoewel simulaties tijd en rekenkracht kosten, helpen ze bij het identificeren van optimale parameter instellingen voor nauwkeurige schattingen in multi-cohort studies met Simon's Two-Stage of andere sequentiële opzetten. Deze bevindingen zijn bedoeld om onderzoekers, vooral in de precisie-oncologie, te ondersteunen bij het maken van geïnformeerde keuzes tijdens de ontwerp- en planningsfase door een goede balans te vinden tussen nauwkeurigheid, consistentie en interpreteerbaarheid.

Een van de belangrijkste conclusies van dit proefschrift is dat schatten van behandeleffecten en onderzoeksopzet nauw met elkaar verbonden zijn in basket trials. Het type ontwerp – single arm, Simon's Two-Stage, of adaptief – bepaalt niet alleen het onderzoek, maar beïnvloedt ook direct de wijze van schatten. Schatten is niet slechts een analyse die volgt op dataverzameling,maar is verweven met het ontwerp zelf, beïnvloed door elementen als steekproefbeperkingen, stopregels, en verwachte variabiliteit tussen cohorten. Dit wordt vooral duidelijk bij schatten onder incomplete inclusie, wat vaak voorkomt bij studies gericht op zeldzame patiëntengroepen. In dergelijke situaties schieten standard schatters zoals de steekproefproportie of zelfs de UMVUE tekort. Zoals aangetoond in dit proefschrift, kunnen Bayesiaanse schatters die informatie delen tussen cohorten stabielere resultaten leveren, hoewel ze sterk afhankelijk zijn van gekozen priors en tuning parameters. Deze parameters zijn niet neutraal; ze weerspiegelen aannames over de verwachte gelijkenis of verschillen in responspercentages tussen baskets. Dit roept een bredere methodologische vraag op: hoe deze aannames duidelijk en transparant te definiëren en verantwoorden, idealiter vóór de start van het onderzoek.

Περίληψη στα Ελληνικά

Στη διατριβή αυτή, διερευνώνται συστηματικά στατιστικές και μεθοδολογικές προσεγγίσεις για την εκτίμηση των θεραπευτικών αποτελεσμάτων σε καινοτόμα σχέδια κλινικών μελετών, με ιδιαίτερη έμφαση στην ογκολογία. Τα τέσσερα κεφάλαια που ακολουθούν περιλαμβάνουν νέες θεωρητικές και πρακτικές μεθοδολογίες, οι οποίες αντιμετωπίζουν συγκεκριμένες προκλήσεις που ανακύπτουν στο σύγχρονο και διαρκώς εξελισσόμενο τοπίο των καινοτόμων σχεδίων κλινικών δοκιμών. Στις συνεισφορές αυτές περιλαμβάνονται: η βελτίωση διαδικασιών εκτίμησης σε περιπτώσεις πρόωρης διακοπής στο κλασικό σχέδιο δύο σταδίων του Simon, η ανάπτυξη μεθόδων εκτίμησης της διάρκειας ανταπόκρισης με βάση λογοκριμένα δεδομένα επιβίωσης, καθώς και η αξιολόγηση Bayesian προσεγγίσεων για τη διαμοίραση πληροφορίας ("borrowing") σε basket trials. Συνολικά, η εργασία αυτή συνεισφέρει στη βαθύτερη κατανόηση του πώς η διαδικασία εκτίμησης αλληλεπιδρά με το σχεδιασμό της μελέτης, και προσφέρει πρακτικές στρατηγικές για πιο αξιόπιστα και ερμηνεύσιμα αποτελέσματα σε σύνθετα ερευνητικά πρωτόκολλα.

Στο **Κεφάλαιο 2**, εξετάζω βελτιωμένες μεθόδους εκτίμησης σε σχέδια STS όταν η μελέτη διακόπτεται πρόωρα. Εστιάζω σε μία ειδική παραλλαγή, διακοπής του σχεδίου δύο σταδίων του Simon (SSTS), όπου η μελέτη σταματά μόλις μπορεί να ληφθεί απόφαση ως προς την αποτελεσματικότητα ή μη της θεραπείας, ακόμη και πριν συμπληρωθεί το αρχικά προβλεπόμενο μέγεθος δείγματος. Αναπτύσσω έναν ομοιόμορφα αμερόληπτο εκτιμητή ελαχίστης διασποράς (UMVUE) για το ποσοστό ανταπόκρισης στη θεραπεία στο SSTS και αποδεικνύω ότι παραμένει αμερόληπτος, παρά τους ευέλικτους κανόνες διακοπής. Επιπλέον, προτείνω μία μέθοδο κατασκευής διαστημάτων εμπιστοσύνης με βάση τη διάταξη του δειγματικού χώρου, η οποία ανταποκρίνεται στον διαδοχικό χαρακτήρα του σχεδίου. Μέσω εκτεταμένων προσομοιώσεων, συγκρίνω τον UMVUE του SSTS με τους συμβατικούς εκτιμητές ως προς τη μεροληψία, το μέσο τετραγωνικό σφάλμα, την πιθανότητα κάλυψης και το αναμενόμενο μέγεθος δείγματος. Τα ευρήματά μου δείχνουν ότι η πρόωρη διακοπή ενδέχεται να μειώσει την ακρίβεια, αλλά ταυτόχρονα οδηγεί σε πιο αποδοτική χρήση χρόνου και πόρων, προσφέροντας δυνητικό όφελος στους ασθενείς.

Στο **Κεφάλαιο 3** μελετάται η εκτίμηση της Περιορισμένης Μέσης Διάρκειας Ανταπόκρισης (Restricted Mean Duration of Response, RMDoR) στη θεραπεία, στον τομέα της ογκολογίας. Η RMDoR προτείνεται ως εναλλακτική της αναμενόμενης διάρκειας ανταπόκρισης (DoR), ειδικά όταν η δεξιά λογοκρισία καθιστά δύσκολη την εκτίμηση της DoR. Η διάρκεια ανταπόκρισης συνήθως βασίζεται σε απεικονιστικά ευρήματα σε προκαθορισμένα χρονικά διαστήματα. Οι διαφορές στα πρωτόκολλα αξιολόγησης μεταξύ μελετών καθιστούν δύσκολη τη σύγκριση των αποτελεσμάτων, κυρίως λόγω διακυμάνσεων στην λογοκρισία διαστημάτων, με άλλα λόγια το πρόγραμμα εξέτασης. Παράλληλα, η διάρκεια παρακολούθησης μπορεί να διαφέρει σημαντικά μεταξύ μελετών, περιπλέκοντας περαιτέρω τις συγκρίσεις και εισάγοντας ενδεχόμενη μεροληψία. Εξετάζω πώς συμπεριφέρεται η RMDoR ως συνάρτηση του χρόνου (τ) και πώς η ερμηνεία της επηρεάζεται από την επιλογή της τιμής αυτής. Προτείνω διαφορετικούς εκτιμητές της RMDoR, λαμβάνοντας υπόψη πρακτικές πτυχές της λογοκρισίας διαστημάτων, είτε χρησιμοποιώντας άμεσα χρόνους ανίχνευσης, είτε με διόρθωση μέσω της εκτίμησης του μέσου διαστήματος (mid-point imputation). Μέσω εκτεταμένων προσομοιώσεων, αξιολογώ την απόδοση αυτών των εκτιμητών τόσο σε δοκιμές με μόνο μία πειραματική θεραπεία όσο και σε τυχαιοποιημένες με-

λέτες, υπό διάφορα σενάρια επιβίωσης και πρωτόκολλα απεικόνισης. Τα αποτελέσματά μου δείχνουν ότι η λογοκρισία διαστημάτων μπορεί να εισάγει μεροληψία στις εκτιμήσεις επιβίωσης, ωστόσο το αποτέλεσμα αυτό είναι περιορισμένο στην RMDoR. Επισημαίνω ότι η επιλογή του τ διαδραματίζει καθοριστικό ρόλο, ιδιαίτερα σε συγκρίσεις μεταξύ θεραπειών. Τονίζω τη σημασία της προσεκτικής αναφοράς των εκτιμήσεων RMDoR ως συνάρτηση του τ και προτείνω τη χρήση του λόγου RMDoR ως ένα σταθερό και ερμηνεύσιμο μέτρο. Τα κριτήρια PASKWIL 2021 [28] αξιολογούν την κλινική σημασία μη τυχαιοποιημένων μελετών βασιζόμενα στο κατώτατο όριο του 95% διαστήματος εμπιστοσύνης του αντικειμενικού ποσοστού ανταπόκρισης (ORR), καθώς και σε σημειακή εκτίμηση της διάμεσης DoR. Το εν λόγω πλαίσιο δεν λαμβάνει υπόψη την αβεβαιότητα της εκτίμησης DoR ούτε το σύνολο του πληθυσμού σε περιπτώσεις σπάνιας ανταπόκρισης. Η συνεισφορά του RMDoR στη σχετική συζήτηση, εξετάζοντας την ως εναλλακτικό μέτρο που περιλαμβάνει τόσο τους ανταποκρινόμενους όσο και τους μη ανταποκρινόμενους στη θεραπεία ασθενείς, εναρμονιζόμενο με την οπτική των estimands που πρότειναν οι Huang et al. [50]. Δείχνω ότι η RMDoR μπορεί να εκτιμηθεί αξιόπιστα, ακόμη και με λογοκρισία διαστημάτων ή δεξιά λογοκρισία, και καταδεικνύω τη μεταβλητότητα της ως προς την τιμή του τ. Για τη χρήση της RMDoR στη λήψη αποφάσεων, συνιστάται ο σαφής και ερμηνεύσιμος καθορισμός ορίων, όπως η μετάφραση των ορίων της DoR των PASKWIL σε ισοδύναμα της RMDoR, με χρήση ιστορικών δεδομένων ή εμπειρογνωμόνων. Η ενσωμάτωση της RMDoR δίπλα στο ORR μπορεί να προσφέρει πληρέστερη εικόνα της αποτελεσματικότητας μιας θεραπείας και να συνεισφέρει στην αναβάθμιση των κλινικών οδηγιών όπως τα κριτήρια PASKWIL, συσχετίζοντας το όφελος με το μέγεθος του αποτελέσματος και την αβεβαιότητα εκτίμησης.

Στο **Κεφάλαιο 4** εστιάζω στην εκτίμηση ποσοστών ανταπόκρισης ανά υπό-μελέτη σε basket trials, όπου μία θεραπεία αξιολογείται σε διαφορετικούς τύπους καρκίνου με κοινό μοριακό στόχο. Τα σχέδια αυτά προσφέρουν λειτουργική αποδοτικότητα και, μέσω Bayesian μεθόδων που επιτρέπουν διαμοίραση πληροφορίας μεταξύ υπό-μελετών, δύνανται να προσδώσουν και στατιστική αποδοτικότητα. Το κεφάλαιο αυτό επικεντρώνεται στην αξιολόγηση της ακρίβειας και της αποδοτικότητας διαφόρων Bayesian εκτιμητών που υλοποιούν αυτή τη διαμοίραση πληροφορίας (“borrowing”), συγκριτικά με τη συμβατική δειγματική εκτίμηση (sample proportion). Μέσω εκτεταμένης προσομοίωσης, συγκρίνω επτά Bayesian εκτιμητές σε ποικίλα ετερογενή σενάρια ποσοστών ανταπόκρισης μεταξύ υπό-μελετών και με διαφορετικά μεγέθη δειγμάτων. Η αξιολόγηση βασίζεται σε τρία κριτήρια: μέση απόλυτη μεροληψία, μέσο τετραγωνικό σφάλμα (MSE) και τον βαθμό συρρίκνωσης προς το συνολικό μέσο ποσοστό ανταπόκρισης. Αποδεικνύω ότι σε περιπτώσεις χαμηλής ετερογένειας, ο εκτιμητής των Berry επιτυγχάνει τη μικρότερη μεροληψία και MSE, λόγω ισχυρής διαμοίρασης πληροφορίας. Σε σενάρια με μεγαλύτερη ετερογένεια μεταξύ υπό-μελετών, κανένας εκτιμητής δεν υπερέχει συστηματικά, η απόδοση εξαρτάται από τα χαρακτηριστικά των δεδομένων, τις εκ των προτέρων παραμέτρους του μοντέλου και τις ιδιαιτερότητες κάθε περίπτωσης. Τα ευρήματα μου αναδεικνύουν το γνωστό δίλημμα μεροληψίας-διασποράς (“bias-variance trade-off”) της διαμοίρασης πληροφορίας: μπορεί να μειώνει τη διασπορά, αλλά και να εισάγει μεροληψία, κυρίως όταν τα πραγματικά ποσοστά ανταπόκρισης διαφέρουν σημαντικά μεταξύ υπό-μελετών. Εξετάζω πώς η επιλογή εκ των προτέρων μεταβλητών (priors), ιδίως για τη παράμετρο του μέσου όρου, μπορεί να επηρεάσει τις εκτιμήσεις και τα συμπεράσματα. Το εύρημα αυτό υπογραμμίζει τη σημασία προσεκτικής επιλογής παραμέτρων στην πράξη. Τονίζω την ανάγκη για επιλογή εκτιμητικής μεθόδου ανάλογα με το πλαίσιο και τη χρήση προσομοιώσεων στον σχεδιασμό, ώστε να καθοδηγείται η επιλογή μοντέλου και παραμέτρων. Οι παρατηρήσεις αυτές δύνανται να υποστηρίξουν τους ερευνητές στον σχεδιασμό basket trials, όπου η ακριβής εκτίμηση ποσοστών ανταπόκρισης είναι κρίσιμη για μελλοντική έρευνα σε ελπιδοφόρες θεραπείες.

Στο **Κεφάλαιο 5** ακολουθώ μια προσέγγιση βασισμένη σε προσομοιώσεις για τη βέλτιστη επιλογή μοντέλου και ρύθμιση παραμέτρων (tuning) κατά τον σχεδιασμό της μελέτης. Για να διασφαλίσω τη συνάφεια με τη μελέτη DRUP, επεκτείνω την αξιολόγηση των ποσοστών ανταπόκρισης σε basket trials που ακολουθούν το σχέδιο δύο σταδίων του Simon, και όχι σε ένα στάδιο όπως στο κεφάλαιο 4, χρησιμοποιώντας χαρακτηριστικά παραδείγματα από τη μελέτη DRUP. Εστιάζω σε τέσσερις Bayesian μεθόδους εκτίμησης (Berry, EXNEX, Psioda και Fujikawa) που επιτρέπουν δια-

μοίραση πληροφορίας μεταξύ υπο-μελετών. Η επιλογή παραμέτρων μέσω προσομοίωσης (“simulation-based tuning”), είναι ιδιαίτερα χρήσιμη στη φάση σχεδιασμού. Εξερευνώντας ένα ευρύ φάσμα τιμών υπερπαραμέτρων, προσδιορίζω βέλτιστες ρυθμίσεις σύμφωνα με δύο κριτήρια χρησιμότητας (“utilities”): την ελαχιστοποίηση της μέσης απόλυτης ρίζας μέσου τετραγωνικού σφάλματος (min-mean aRMSE) και την ελαχιστοποίηση του σφάλματος στο δυσμενέστερο σενάριο (min-max aRMSE). Τα κριτήρια αυτά αντανακλούν διαφορετικές προτεραιότητες: ακριβή εκτίμηση κατά μέσο όρο έναντι αποφυγής ακραίων λαθών. Προσομοιώνω επίσης σενάριο με περιορισμένο μέγεθος δείγματος, αν η δεύτερη φάση μιας υπο-μελέτης δεν μπορεί να συμπεριληφθεί πλήρως στην εκτίμηση του ποσοστού ανταπόκρισης. Τα αποτελέσματα καταδεικνύουν ότι, με βάση τα παραπάνω κριτήρια, και οι τέσσερις Bayesian μέθοδοι μπορούν να υπερέχουν του δειγματικου εκτιμητή, ειδικά σε περιπτώσεις ελλιπούς συμμετοχής. Για όλες τις μεθόδους εντοπίστηκαν παραμετρικές ρυθμίσεις με εξαιρετική επίδοση ως προς το aRMSE. Οι μέθοδοι Berry, Fujikawa και Psioda έδειξαν ευαισθησία στην επιλογή priors σε ορισμένα σενάρια, γεγονός που υπογραμμίζει τη σημασία προσεκτικού tuning. Η μέθοδος EXNEX αποδείχθηκε η πιο σταθερή σε όλο το φάσμα παραμέτρων και σεναρίων, αν και το εύρος διερεύνησης παραμέτρων ήταν πιο περιορισμένο σε σύγκριση με τις υπόλοιπες. Εφαρμόζοντας τις μεθόδους αυτές σε πραγματικά δεδομένα DRUP, αποδεικνύω πώς διαφορετικοί εκτιμητές οδηγούν σε διαφορετικά αποτελέσματα ανά μελέτη, γεγονός με άμεσες επιπτώσεις στην κλινική ερμηνεία. Σε υπο-μελέτες με υψηλό παρατηρούμενο ποσοστό ανταπόκρισης, η διαμοίραση πληροφορίας ενδέχεται να οδηγήσει σε χαμηλότερες εκτιμήσεις, λόγω της συρρίκνωσης προς τον μέσο, ενώ σε μικρές ή αβέβαιες υπο-μελέτες το αποτέλεσμα μπορεί να είναι το αντίστροφο. Η δυναμική αυτή αναδεικνύει την αναγκαιότητα προκαθορισμού της εκτιμητικής μεθόδου στο ερευνητικό πρωτόκολλο. Το κεφάλαιο καταδεικνύει ότι η Bayesian διαμοίραση πληροφορίας δεν αποτελεί πανάκεια, αλλά μια ευέλικτη στρατηγική που απαιτεί προσεκτική επιλογή priors και tuning παραμέτρων. Αν και οι προσομοιώσεις απαιτούν χρόνο και υπολογιστική ισχύ, συμβάλλουν στον εντοπισμό βέλτιστων ρυθμίσεων για ακριβείς εκτιμήσεις σε μελέτες με πολλαπλές υπο-μελέτες και σχέδια δύο σταδίων. Τα συμπεράσματα αυτά στοχεύουν στην υποστήριξη των ερευνητών, κυρίως στην εξατομικευμένη ογκολογία, κατά τη λήψη τεκμηριωμένων αποφάσεων στο σχεδιασμό και την προγραμματισμό μελετών, με έμφαση στην επίτευξη ισορροπίας μεταξύ ακρίβειας, συνέπειας και ερμηνευσιμότητας.

Ένα από τα σημαντικότερα συμπεράσματα της διατριβής είναι ότι η εκτίμηση των θεραπευτικών αποτελεσμάτων και ο σχεδιασμός της μελέτης είναι στενά συνδεδεμένα στα basket trials. Ο τύπος του σχεδίου, είτε πρόκειται για δοκιμή ενός σταδίου, είτε για σχέδιο δύο σταδίων του Simon ή για προσαρμοστικό σχέδιο, δεν καθορίζει μόνο τον τρόπο διεξαγωγής της μελέτης, αλλά επηρεάζει άμεσα και τον τρόπο εκτίμησης των αποτελεσμάτων. Η εκτίμηση δεν αποτελεί απλώς μια στατιστική ανάλυση μετά τη συλλογή των δεδομένων, αλλά επηρεάζει άμεσα με τον ίδιο το σχεδιασμό της μελέτης, υπό την επίδραση παραγόντων όπως οι περιορισμοί του δείγματος, οι κανόνες διακοπής και η αναμενόμενη μεταβλητότητα μεταξύ υπο-μελετών. Αυτό γίνεται ιδιαίτερα εμφανές σε περιπτώσεις ελλιπούς συμμετοχής, που συχνά παρατηρούνται σε μελέτες με ασθενών με σπάνιες παθήσεις. Σε τέτοιες περιπτώσεις, οι κλασικοί εκτιμητές, όπως η δειγματική αναλογία, συχνά δεν επαρκεί. Όπως καταδεικνύεται στη διατριβή, οι Bayesian εκτιμητές που επιτρέπουν διαμοίραση πληροφορίας μεταξύ υπο-μελετών μπορούν να προσφέρουν σταθερότερα αποτελέσματα, αν και εξαρτώνται έντονα από την επιλογή των priors και των παραμέτρων tuning. Αυτές οι παράμετροι δεν είναι ουδέτερες, αντανακλούν υποθέσεις σχετικά με την αναμενόμενη ομοιότητα ή διαφορά στα ποσοστά ανταπόκρισης μεταξύ baskets. Το γεγονός αυτό αναδεικνύει ένα ευρύτερο μεθοδολογικό ερώτημα: πώς οι υποθέσεις αυτές μπορούν να οριστούν και να αιτιολογηθούν με σαφήνεια και διαφάνεια, ιδανικά πριν από την έναρξη της μελέτης.

Acknowledgements

The PhD studies are a marathon, not a sprint. Looking back on this process, full of knowledge, difficulties and joy, I feel blessed to have met people who joined me and supported me in this long run. I would like to acknowledge the support of these people who helped me become a better person and finally complete my PhD.

Prof. dr. K.C.B. Roes, dear Kit, I feel profoundly fortunate to have you as my promotor. Thank you for being there from day one, for your steady guidance, the honest critiques when I was drifting from our target, and your calm leadership when things got hard. You taught me to slow down, think clearly, and trust the process. I grew not only in statistics but as a person: more patient, more careful with the details, and more confident in making choices. I will never forget the conversation we had after the first year of my PhD. I'm deeply grateful for your time, your patience, and the many conversations that kept me on track.

Dr. V. van der Noor, dear Vincent, thank you for your kindness and the quiet confidence you showed in me. Your curiosity and love for research were contagious, they kept me going when my energy ran low. Your mentoring felt like an invitation to learn, not a test to pass. I didn't realise how powerful a few encouraging words could be until I read your comments. Thank you for believing in this project and for answering every question, big or small.

Dr. R.M. van den Bor, dear Rutger, I'm grateful that our paths crossed—professionally and personally. As my daily supervisor, you were there in the thick of the work, helping me carry this project forward, task by task. Your insistence on detail, our methodological debates, and the many day-to-day discussions shaped me into a more rigorous researcher and a more grounded person more mature and better able to handle difficult moments. Thank you for standing by this project, even when your schedule was overflowing, and for believing in it (and in me) all the way through.

Dr. M.A. Jonker, dear Marianne, thank you for stepping in at a crucial moment and for the clarity you brought to the hardest methodological parts of this thesis. Your analytical way of thinking, your careful guidance, and your steady encouragement helped me find my footing. You taught me to be precise and to give attention to the small details that make the work strong. I am deeply grateful for your time, your patience, and your faith in this project.

To the members of the assessment committee and the opponents, I am honored by the care and time you devoted to reading my thesis and engaging with it so thoughtfully.

Dr. H. van Tinteren, dear Harm, thank you for giving me the opportunity to start this project and begin my career at the NKI. I'm deeply grateful for your support in my first months in the Netherlands and for your guidance

as we shaped the research questions. The clinical research experience I gained as a consultant at the institute has been invaluable. I will always remember your calm, clear way of explaining things and the lasting impact you had on this project and on me.

I am deeply thankful to the Radboud UMC for embracing this project and supporting it.

Thank to the Julius centrum for hosting me in the first years of my PhD project. Your welcoming environment, practical support, and day-to-day conversations helped me find my footing and grow.

To my NKI colleagues in the WA, and to my statistician colleagues, Erik, Karolina, Rob, and Mutamba, thank you for your support and generous help throughout these years. Your questions sharpened my thinking, and your willingness to discuss, share feedback, and troubleshoot when it mattered made a difference.

Dear Sara, Marta and Shermarke, I want to thank you for the your support throughout this time. Each of you supported me in your own way, always thoughtful, always encouraging, and your positivity kept me going when things were tough.

Dear Bert, I would like to thank you for the opportunity you gave me to finish my PhD project, I couldn't do this without your help. Your leadership, clear guidance and support allowed me to become a better professional and those lessons have carried into many parts of my life.

Dear Marian, thank you for your help and support in this journey. I couldn't make it without your support in the beginning of the project. Your kindness, patience, and willingness to help taught me so much about the work and about how to show up for others. I'm grateful for our friendship.

Dear Adrianna, I'm so grateful you were part of this. I still think of our coffee breaks and the patience you showed with my endless questions, never once making me feel they were that stupid. You helped me grow as a statistician and, even more, as a friend. Thank you for the encouragement, the practical tips, and the time we spend discussing about everything all these years.

I could not have completed this PhD without the support of my friends. I want to thank, Stelios, Nikitas, Manolis, Fotis, Alex, Nikos, Eirini, Zena, Anna-Maria, Tsaltas, Tsepas, Kanakis, Georgia and Alex, Krou, Sifis, Thomas, Giorgos Sei, Lefteris, Thodoris and Konstantinos for showing up in big and small ways. After all these years that the code was running, finally I made it. A special thank also to Bill, Nikiforos, Andreas, Rafa, Nikos and Kostas for their support. I want also to thank for the support of Katerina, Theo and Iro, you have been vital in encouraging me.

Dear Gianni, koumparouli, thanks for the support, thank you for listening to me in my happy or even in my complaining ones.

Dear Lelio and Maria, I can't overstate how important your support has been. Even from afar, I always felt you by my side. You helped me keep perspective and keep going. I truly consider you part of this success. Thank you for making the distance feel small.

Dear George, Dimitri and Stefi, from day one to this final stretch, you've been there. There were many turning points in this process that I cannot speak of how crucial was your support, a video call or just an evening at home. I feel we grew up all together in this. An honorable mention goes for little Nikolas.

Στην οικογένεια μου, στις γιαγιάδες μου Ελένη και Κωνσταντίνα, σε όλους μου τους θείους και τις θείες, τα ξαδέρφια μου ευψαριστώ πολύ για τη στήριξή σας όλα αυτά τα χρόνια, τις όμορφες στιγμές και την αγάπη που μου δείξατε.

Μάνα και Πατέρα, ευχαριστώ πολύ για όλα όσα μου έχετε δώσει. Δε θα μπορούσα να ζητήσω περσσότερα και να καταφέρω τίποτα χωρίς την αγάπη σας, τη βοήθεια σας, τη καθοδήγηση και το τρόπο σας.

Αδερφέ μου, όλη αυτή η πορεία μας έφερε πιο κοντά, και είμαι ευγνωμον για αυτό. Ευχαριστώ πολύ για όλη τη στηριξή σου σε κάθε στιγμή εύκολη ή δυσκολή.

Χριστιάννα μου, σε ευχαριστώ για όλη την υπομονή, την υποστήριξη, την κατανόηση σου σε όλη αυτή τη πορεία, στις δύσκολες στιγμές, και στις όμορφες, μου δίνεις δύναμη να προσπαθώ όσο γίνεται περισσότερο στη κάθε μέρα μου.

About the Author

Antonios Daletzakis was born on 24 May 1992 in Athens, Greece. After graduating from the 3rd General Lyceum of Agioi Anargyroi, Attica in 2010, he pursued undergraduate studies in Mathematics at the University of Crete, earning a BSc in 2015. In 2016, he completed an MSc in Applied Statistics at the Athens University of Economics and Business, developing broad expertise across statistical methodology and conducting a master's thesis using real trial data.

Following graduation, he fulfilled mandatory military service in the Hellenic Army on the island of Rhodes (2017). In 2018, he began doctoral studies at the Netherlands Cancer Institute (NKI) in collaboration with the Julius Center, UMC Utrecht. As part of the institute's consultancy group, he gained hands-on experience in clinical trial design, protocol development, and statistical analysis of clinical data, working alongside clinical researchers and statisticians. His doctoral research focuses on methods for estimation in clinical trials in two-stage trials, basket trials and the development of a Duration of response estimation. In November 2022, he began working full time as a Statistician in the Biometrics department at NKI, with the doctoral appointment subsequently transferred to Radboud University Medical Center (Radboudumc).

List of Publications

1. Chantal F. Stockem, Jeroen van Dorp, Nick van Dijk, Daniel J. Vis, Rolf Harkes, Bram van den Broek, Maartje Alkemade, Annegien Broeks, Kees Hendricksen, Thierry N. Boellaard, Jeantine M. de Feijter, Maurits L. van Montfoort, **Antonios Daletzakis**, Antoine G. van der Heijden, Richard P. Meijer, Niven Mehra, Lodewyk F.A. Wessels, Bas W.G. van Rhijn, Britt B.M. Suelmann, Michiel S. van der Heijden, "Final clinical analysis of pre-operative ipilimumab and nivolumab in locally advanced urothelial cancer and exploration of tumor-draining lymph node composition: The NABUCCO trial," *European Journal of Cancer*, vol. 229, 115731, 2025.

2. **Antonios Daletzakis**, Kit C. B. Roes, Marianne A. Jonker, "Estimation of the Restricted Mean Duration of Response (RMDoR) in Oncology," *Pharmaceutical Statistics*, vol. 24, no. 1, e2468, 2025.

3. A. van Ommen-Nijhof, T.G. Steenbruggen, T.G. Wiersma, S. Balduzzi, **A. Daletzakis**, M.J. Holtkamp, M. Delfos, M. Schot, K. Beelen, E.J.M. Siemerink, J. Heijns, I.A. Mandjes, J. Wesseling, E.H. Rosenberg, M.J.T. Vrancken Peeters, S.C. Linn, G.S. Sonke, "Intensified alkylating chemotherapy for patients with oligometastatic breast cancer harboring homologous recombination deficiency: Primary outcomes from the randomized phase III OLIGO study," *European Journal of Cancer*, vol. 213, 115083, 2024.

4. Bianca A. M. H. van Veggel, Anthonie J. van der Wekken, Marthe S. Paats, Lizza E. L. Hendriks, Sayed M. S. Hashemi, **Antonios Daletzakis**, Daan van den Broek, Linda J. W. Bosch, Kim Monkhorst, Egbert F. Smit, Adrianus J. de Langen, "A phase 2 trial combining afatinib with cetuximab in patients with EGFR exon 20 insertion–positive non–small cell lung cancer," *Cancer*, vol. 130, no. 5, pp. 683–691, 2024.

5. Jeroen van Dorp, Christodoulos Pipinikas, Britt B. M. Suelmann, Niven Mehra, Nick van Dijk, Giovanni Marsico, Maurits L. van Montfoort, Sophie Hackinger, Linde M. Braaf, Tauanne Amarante, Charlaine van Steenis, Kirsten McLay, **Antonios Daletzakis**, Daan van den Broek, Maaike W. van de Kamp, Kees Hendricksen, Jeantine M. de Feijter, Thierry N. Boellaard, Richard P. Meijer, Antoine G. van der Heijden, Nitzan Rosenfeld, Bas W. G. van Rhijn, Greg Jones, Michiel S. van der Heijden, "High- or low-dose preoperative ipilimumab plus nivolumab in stage III urothelial cancer: the phase 1B NABUCCO trial," *Nature Medicine*, vol. 29, no. 3, pp. 588–592, 2023.

6. **Antonios Daletzakis**, Rutger van den Bor, Marianne A. Jonker, Kit C. B. Roes, Harm van Tinteren, "Estimation

and expected sample size in Simon's two-stage designs that stop as early as possible," *Pharmaceutical Statistics*, vol. 21, no. 5, pp. 879–894, 2022.

7. Leonora de Boo, Ashley Cimino-Mathews, Yoni Lubeck, **Antonios Daletzakis**, Mark Opdam, Joyce Sanders, Erik Hooijberg, Annelot van Rossum, Zuzana Loncova, Dietmar Rieder, Zlatko Trajanoski, Marieke Vollebergh, Marcelo Sobral-Leite, Koen van de Vijver, Annegien Broeks, Rianne van der Wiel, Harm van Tinteren, Sabine Linn, Hugo Mark Horlings, Marleen Kok, "Tumour-infiltrating lymphocytes (TILs) and BRCA-like status in stage III breast cancer patients randomised to adjuvant intensified platinum-based chemotherapy versus conventional chemotherapy," *European Journal of Cancer*, vol. 127, pp. 240–250, 2020.

Radboudumc
university medical center
IN DEI NOMINE FELICITER
Radboud University