Design and Analysis of Basket Clinical Trials

Elizabeth Daniells, MSci (Hons), MRes

Submitted for the degree of Doctor of Philosophy at Lancaster University.

September 2024

STOR-i
excellence with impact

Abstract

The rise of precision medicine has ushered in the development of innovative clinical trial designs, notably basket trials, which assess the efficacy of a single therapeutic treatment across multiple disease types simultaneously, with each disease group forming a 'basket'. Basket trials are advantageous as they allow testing of treatments on rare disease types which do not typically warrant their own investigation due to limited sample sizes. Small sample sizes can result in a lack of statistical power and precision of estimates. Bayesian information borrowing models can be implemented to improve inference by leveraging information from one basket when making inference in another. This thesis develops novel information borrowing methodology to improve power and reduce the type I error rate under various settings.

This thesis first explores several existing Bayesian information borrowing models and proposes a novel data-driven adaptation. The models are investigated through simulation studies under numerous settings, including the often-overlooked unequal sample size case. Results indicate that the proposed approach better controls for a type I error, whilst yielding improved power.

Approaches for the addition of new baskets to an ongoing trial are also proposed. Our findings demonstrate a substantial improvement in power in new baskets when information borrowing is utilised, though this comes with the risk of error inflation. We propose a novel calibration of efficacy criteria to mitigate this inflation. Simulation results show that implementing this calibration reduces error rates, with only a small

loss in power in a few cases.

Within the literature there are typically two avenues for information borrowing: borrowing between baskets on a trial or borrowing from historic data. We develop models that amalgamate both forms of borrowing. We show that the incorporation of historic data can improve power of estimates, whilst maintaining similar error rates to a method that ignores historic data.

Dedicated to the village who raised me.

"Put a hand on your heart, say whatever you feel, be wherever you are"

- Noah Kahan

Acknowledgements

This PhD has been a roller coaster of a journey and it would have not been possible without the support of so many wonderful people. I would first like to thank my academic supervisors Pavel Mozgunov and Thomas Jaki who have guided me through this PhD, and made the seemingly 'unachievable' goal of completing a PhD not only achievable, but an enjoyable process. You have taught me so much in the past 4 years, your guidance and expertise have been instrumental in shaping me into an efficient and effective researcher. You have also helped me thrive as a human being, building my confidence and pushing me to my potential. I am extremely grateful for the opportunities that you provide me with and look forward to continuing my career under your mentorship. I would also like to thank my Lancaster supervisor, Helen Barnett who joined late on in the project but provided me valuable life advice and space to vent each week. I also had the privilege of working in partnership with Alun Bedding from Roche, whose enthusiasm for my research shone through at each meeting.

I am so very grateful that I have completed this PhD at the STOR-i Centre for Doctoral Training and I would like to thank all of the staff for their unwavering support and for believing in me when I did not believe in myself. In particular, I would like to thank Wendy and Jon for getting me through some of my lowest moments and celebrating my wins, I would not be here today without you both. Thanks also to Nicky, Kim, Idris, Anna and Rachel for all the support and opportunities you provided throughout my time in STOR-i. A special thanks to all the friends I have made in the

centre over the years.

I would also like to thank my maths teachers from secondary school, Mrs Burlace and Mrs Barker who were instrumental in sparking my love for maths and for getting me to university in the first place. You are fantastic role models who made being a female in mathematics a norm for me.

Lancaster University has been my home for the past 9 years. I have gained a family for life at LU Trampolining Club. Being part of this close knit team has shaped the person I am today and has taught me the value of true friendship and togetherness. I will always cherish the memories we have made and I cannot wait to see what you all achieve in the future.

I have been lucky to have had many cheerleaders in the wider Lancaster community, so I would also like to acknowledge Josie, Jess and Rachel, as well as, Holly and Leigh from Red Rose (thank you for listening to all of my ranting phone calls and the random conversations when I needed them most).

Most importantly, I would like to thank my family, this PhD is dedicated to you all. To my parents for encouraging me in every endeavour, for loving and supporting me through thick and thin and always believing in my potential. Thank you Dad for sending me the link to STOR-i's website during my second year of undergrad, you were right, it was perfect for me! To Emily and Elsie, the best sisters I could ask for, keeping me entertained whenever I came home and calling me out on my excessive comma usage (hopefully there are not too many in this thesis). To Grandma and Grandad, I simply do not have words to sum up what you mean to me but I love you more than anything and I know I have made you proud. And finally to Grandad Gower, Sue, all my aunts, uncles and cousins, thank you for being ever-present in my life and cheering me on to the end of this PhD. A special shout-out to Granny, I know she would have been with me every step of the way.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 2 has been published as: Daniells, L., Mozgunov, P., Bedding, A., Jaki, T. A comparison of Bayesian information borrowing methods in basket trials and a novel proposal of modified exchangeability-nonexchangeability method. Stat Med. 2023;42(24):4392-4417. doi:10.1002/sim.9867

Chapter 3 is currently under review in a peer-reviewed journal and has been submitted for publication as: Daniells, L., Mozgunov, P., Bedding, A., Barnett, H., Jaki, T. How to add baskets to an ongoing basket trial with information borrowing.

Chapter 4 is currently under review in a peer-reviewed journal and has been submitted for publication as: Daniells, L., Mozgunov, P., Bedding, Barnett, H., A., Jaki, T. Incorporating historic information to further improve power when conducting Bayesian information borrowing in basket trials.

The word count for this thesis is approximately 46,000 words.

Elizabeth Daniells

Contents

List of Figures

List of Tables

List of Abbreviations

BaCIS	Bayesian hierarchical classification and information sharing
BCHM	Bayesian Cluster Hierarchical Model
BHM	Bayesian Hierarchical Model
BMA	Bayesian Model Averaging
CBHM	Calibrated Bayesian Hierarchical Model
CPP	Commensurate Predictive Prior
ECD	Erdheim-Chester Disease
EX	Exchangeable
EXNEX	Exchangeability-Nonexchangeability
$\mathbf{EXNEX_{pool}}$	Exchangeability-Nonexchangeability with pooled historic and current data
EXppNEX	Exchangeability-Nonexchangeability with power prior in the NEX component
FDA	US Food and Drug Administration
FWER	Family-Wise Error Rate
histFujikawa	Adapted Fujikawa's design incorporating historic data
IND	Independent analysis of new baskets added to a trial
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
LCH	Langerhans'-Cell Histiocytosis
MAP	Meta-Analytic-Predictive Prior
$\mathbf{mEXNEX}_c$	Modified Exchangeability-Nonexchangeability model with cut-off c

mEXNEX$_{\text{hist}}$	Modified Exchangeability-Nonexchangeability model with historic data
MLMixture	Multi-Level Mixture model
MPP	Modified Power Prior
MUCE	Multiple Cohort Expansion
NEX	Nonexchangeable
NSCLC	Non-Small-Cell Lung Cancer
ORR	Overall Response Rate
PL1	Planned addition of new baskets with a single EXNEX model applied
PL2	Planned addition of new baskets with two EXNEX models applied
PP	Power Prior
RCaP	Robust Calibration Procedure
RCT	Randomized Controlled Trial
RoBoT	Robust Bayesian Hypothesis Testing
SAM	Self-Adapting Mixture Prior
UNPL	Unplanned addition of new baskets to a trial
% Reject	% data sets in which the null hypothesis is rejected
% All Correct	% data sets in which the correct conclusions are made in all baskets

List of Symbols

α	Significance level *or* power parameter in a power prior.
q_0	Null response rate.
q_1	Target response rate.
q_2	Marginally effective response rate.
H_0	Null hypothesis.
H_1	Alternative hypothesis.
D	Observed response data.
$\mathbb{P}(\cdot)$	Probability.
$\mathbb{I}(\cdot)$	Indicator function.
$L(\cdot\|\cdot)$	Likelihood.
$\pi_0(\cdot)$	Prior distribution.
K	Total number of baskets.
k	Denotes a basket $1, \ldots, K$.
Y_k	Number of responses in basket k.
n_k	Sample size in basket k.
p_k	Response rate in basket k.
p	True response rate vector.
$\hat{p}_k$	Point estimate of the response rate in basket k.
Δ_α	Efficacy cut-off for posterior probabilities, controlled at the $\alpha\%$ level.
Δ_k	Efficacy cut-off for posterior probabilities of basket k.

θ_k	Logit-transformed response rate, p_k, for basket k.
μ	Common mean parameter for the normal prior on θ_k in the hierarchical models.
σ^2	Common variance parameter for the normal prior on θ_k in the hierarchical models.
m_μ	Hyper-prior mean parameter for μ in the hierarchical models.
ν_μ	Hyper-prior variance parameter for μ in the hierarchical models.
m_k	NEX mean parameter in the EXNEX model.
ν_k	NEX variance parameter in the EXNEX model.
π_k	Prior probability of exchangeability in basket k.
δ_k	Binary mixture weight for basket k in the EXNEX model.
ρ_k	A plausible guess for the true response rate p_k.
T	Chi-squared test statistic of homogeneity.
a, b	Tuning parameters in the CBHM.
c	Cut-off value in the mEXNEX$_c$ model, indicating no borrowing should occur.
$h_{k,k'}$	Hellinger distance between posteriors of basket k and k'.
S	Set of all baskets not excluded for heterogeneity in the mEXNEX$_c$ model.
$\mathcal{M}_j$	Model j representing basket allocations to the EX/NEX groups.
P_j	Number of distinct response rates in model $\mathcal{M}_j$.
k'	Denotes a new basket added to a trial.
K'	Total number of new baskets added to a trial.
k_0	Denotes an existing basket that started a trial.
K_0	Total number of existing baskets that started a trial.
M	Total number of scenarios input into the RCaP.
$\boldsymbol{p_m}$	Response rate for simulation scenario, m, to be input into the RCaP.
$\boldsymbol{n_m}$	Sample size for simulation scenario, m, to be input into the RCaP.
ω_m	Weight for scenario m in the RCaP.
Q	A quantity obtained from the posterior distribution.

$\boldsymbol{Q_k}$	Vectors in the RCaP consisting of values of Q.
H_k	The total number of historic sources of data for basket k.
$k^{*(j)}$	Denotes a basket in historic study j corresponding to current basket k.
k^*	Denotes a historic basket corresponding to current basket k when there is a single source of historic data.
Y_{k^*}	Number of responses in historic basket k^*.
n_{k^*}	Sample size in historic basket k^*.
p_{k^*}	Response rate in historic basket k^*.
α_j	The power parameter in the power prior for historic study j.
ψ_i	Indicator if basket i is a historic basket.
$\pi_{\lambda,k}$	Mixture weights between two EXNEX models in the MLMixture model.
$\pi_{\text{all},i}, \pi_{\text{curr,i}}$	Mixture weights within EXNEX models in the MLMixture model.
$\omega_{k,i}$	Weights in Fujikawa's design from JSD between baskets k and i.
ϵ, τ	Turning parameters in Fujikawa's design.

Chapter 1

Introduction

1.1 The Drug Development Process

In a world prevalent with disease, there is a continual need for the development of new treatments for human health conditions. These treatments can vary from diets and therapies to drugs and surgeries, however, their development is not an easy one. Any new remedies must undergo rigorous testing in order to assess safety, efficacy and dosage before they are made available to the general population. This is a lengthy and costly process. Although it varies based on therapeutic area, Turner (2010) states that drug development could take anywhere from 10 to 15 years and, according to Rosier et al. (2014), cost anywhere from 800 million to 1 billion US dollars.

The US food and drug administration (FDA) and other regulatory authorities impose strict regulations on the drug development process in order to protect public health (Fleming et al., 2017), balancing provision of effective treatments to those in need and minimising potential dangers. The FDA split the drug development process into several stages beginning with discovery and development in which potential candidate treatments are identified and the molecular compounds tested. The second stage is pre-clinical development, where the treatment is tested on animals to identify the

pharmacokinetic profile (Steinmetz and Spack, 2009) and to screen for safety issues (Bravo et al., 2022). Should treatments be deemed safe for human testing, the next stage is clinical research in which the treatment is tested on humans via clinical trials. These clinical trials themselves consist of several phases: identifying safety concerns, exploring dosages and comparing treatments to a placebo or current standard therapy in order to assess efficacy. A success at this stage leads to review by the regulatory authorities for approval for administration to the general population.

Due to the strict nature of this process, only 10-20% of treatments that entered the clinical phase are eventually approved after the regulatory review. According to Yamaguchi et al. (2021), this approval rate has not changed in the past few decades despite advances in science. This motivates the need for improvement in clinical trial design and analysis in order to speed up the process and get the right treatments to the right patients in a more efficient manner. Thorough research into efficient clinical trials has led to advancement in trial designs that expedite the drug development process, however, there is still a long way to go, particularly when it comes to implementing such efficient trial designs in practice.

1.2 Clinical Trials

Whilst the concept of clinical trials dates back to "The Book of Daniel" in the Bible, the first recorded controlled clinical trial was conducted by James Lind in 1747 who conducted a comparative trial for several treatments for scurvy, a prominent issue for sailors on board the HMS Salisbury at sea (Bhatt, 2010). The conclusion of the trial was that oranges and lemons provided the best outcomes, a finding that lines up with what we know today to be the cause of scurvy: a lack of vitamin C. This milestone in the history of clinical trials was followed up by the first placebo-controlled trial in 1863, where a dummy treatment (placebo) was compared to a herbal extract for the

treatment of patients suffering with rheumatism.

Randomisation was first used by Fisher (1926) in the design and analysis of experiments, focusing on application to the agricultural setting. However, randomisation did not become common practice in clinical trials until post World War II. A Randomised controlled trial (RCT) involves randomising patients onto one of two or more treatment arms. Patients on the control arm receive either the current standard treatment or a placebo, while the rest are given new experimental treatments. The purpose of randomisation is to create a balance of patient characteristics across treatment arms (Hariton and Locascio, 2018). Without this balance, bias may be introduced as the differences between the treatment effects cannot be distinguished from the effect driven by differences in patient characteristics. The first RCT was conducted by Sir Bradford Hill with the medical research council in 1946, studying streptomycin on patients with pulmonary tuberculosis (Crofton, 2006). This ground-breaking work continues to influence the field of clinical trial design and analysis to this date. That being said, according to Jones and Podolsky (2015), the RCT was not actually stated as gold-standard until 1982 (Feinstein and Horwitz, 1982). However, back in 1962, the FDA did mandate a "well-controlled" study to demonstrate efficacy as an update to the federal food, drug and cosmetic act (Meadows, 2006).

Although RCTs are still considered the gold standard to this date - promoting rigorous testing - they do come with limitations. The biggest concern comes back to their cost and duration, but they also pose ethical concerns. For instance, it is not considered ethical to set up a trial in which evidence suggests that patients in the control arm of the study will likely to benefit from the experimental treatment, yet still receive a placebo (Stolberg et al., 2004). To add to this, RCTs often require large sample sizes in order to assess efficacy, something infeasible for trials considering rare diseases or conditions. These deficiencies promote the investigation of more advanced trial designs, with the use of 'adaptive' clinical trials significantly increasing in popularity

and implementation in recent years. In fact, the FDA published guidelines in 2019 for the implementation of these adaptive designs in clinical trials (Kaizer et al., 2023a).

Modern clinical trials tend to involve a treatment passing through the following four stages (Sedgwick, 2011):

- **Phase I** - a small trial consisting of a handful of volunteers. Its goal is to test the safety of the treatment, observe major side effects and examine how the body interacts with the treatment (pharmacokinetics),
- **Phase II** - involves a larger group of patients who often suffer from the targeted condition. It concerns finding a safe dose and observes the effectiveness of the treatment,
- **Phase III** - compares the new experimental treatment to the current standard therapy or a placebo. This stage can involve hundreds to thousands of patients,
- **Phase IV** - observes the long term effect and rare side effects of the treatment once it has been released to the market This stage can also involve thousands of patients.

Adaptive trial designs provide more flexibility compared to RCTs, with some allowing a treatment to move between these phases under a single protocol (Lang, 2011). This efficient design significantly decreases the duration of a trial and allows for changes to the study based on analysis conducted at interim time points.

FDA guidelines were published in December 2023 for the use of master protocols in drug development (U.S Food and Drug Administration, 2023b). Master protocols utilise a single overarching trial protocol to address multiple hypotheses simultaneously. This can consist of multiple disease types under investigation, multiple treatments or both (Woodcock and LaVange, 2017). Master protocols share trial procedures including patient enrolment/selection, analysis and data management, improving the efficiency of the study (Hirakawa et al., 2018).

There are three main branches of master protocol: platform trials, umbrella trials, and as is the focus of this thesis, basket trials. Platform trials allow multiple treatments to be evaluated for a single disease simultaneously, whilst allowing treatments to be added or dropped in an adaptive manner. Platform trials played a vital role in the global effort to tackle the COVID-19 pandemic, rapidly identifying effective treatments. The most well known of these COVID-19 platform trials was the RECOVERY trial which was conducted in the United Kingdom. The RECOVERY trial launched in early 2020 with the emergence of the virus and identified several effective treatments, one of which was identified within 100 days of the trial opening (Pessoa-Amorim et al., 2021). This study redefined the potential of streamlined and efficient clinical trial design, which is further supported by the regulatory guidance provided by the FDA in 2023.

1.3 Personalised Medicine & Master Protocols

All three of the main branches of master protocol are in the realm of personalised medicine. Personalised medicine refers to treatments that are targeted to an individuals specific characteristics, as opposed to a disease type on a whole. This tailors treatments to a patients intrinsic factors such as genetic make-up, environmental exposures and lifestyle choices. The need for personalised medicine arises due to individual variability between patients suffering from the same condition, resulting in inevitable heterogeneity in patients responses to treatments (Goetz and Schork, 2018). Therefore, the key question in personalised medicine trials is whether the treatment works uniformly across all patients, across all disease sub-types or are responses patient specific?

The feasibility of a personalised approach has improved with recent advancements in genetic screening and diagnostic techniques. Differences between patient characteristics are typically detected through biomarkers, which are measurable indicators of biological conditions/processes. These biomarkers are often used to identify target populations

within clinical trials (Atkinson et al., 2001).

Biomarkers play a key role in selecting patient populations in the previously discussed platform trials, whilst forming the core rational for the two other branches of master protocol, umbrella and basket trials. Umbrella trials refer to studies in which multiple treatments are tested in parallel on patients who share a single disease type but are characterised by different biomarkers (Di Liello et al., 2021). In contrast, basket trials test a single treatment on multiple patient populations suffering from different conditions, where patients form sub-groups (also called 'baskets') based on their condition. Patients within each basket share a common biomarker (Park et al., 2019). Whilst these trials have been implemented in the full range of drug development process, through phases I-IV, they are most common in early phase trials in which determining efficacy is the primary objective (Ouma et al., 2022b).

Personalised medicine is fundamental in the field of oncology. Sargent and Renfro (2017) discuss how, in previous designs, all cancers were considered as homogeneous and thus all received the same treatment. However, with the development and inclusion of biomarkers in study designs, it is now possible to stratify patients into separate treatment groups based on their biomarker status. Each biomarker sub-group could be considered in independent two-arm studies, each with their own study protocol. This is extremely time consuming, ineffective and requires a large trial infrastructure.

Master protocols are key to clinical trials in a personalised medicine setting as they are designed to test a treatment on small groups of patients in parallel. However, as discussed by Strzebonska and Waligora (2019), some limitations still remain. For example, some may question the scientific validity of results based on biomarker stratification in the case of umbrella trials. In practice, patients will likely present one or more of the biomarkers in question, therefore the question surrounds how they are then allocated to biomarker group and hence a treatment. Should this issue not be addressed, clinicians may induce bias by favouring one treatment over another. Strzebonska and Waligora

(2019) also discuss the use of surrogate endpoints (substitutes for clinically meaningful outcomes) in basket and umbrella trials, where the focus is often overall survival or tumour shrinkage. According to Kemp and Prasad (2017), 66% of oncology trials between 2009 and 2014 were approved based on surrogate endpoints. The advantage of this is that their outcome is observed a lot quicker than the long-term outcome, speeding up the trial process. However, Kemp and Prasad (2017) claim that the use of surrogate endpoints may fall short when there is a level of uncertainty in their correlation to the original endpoint of interest. A final disadvantage in master protocols and precision medicine in particular is the issue of informed consent. Throughout recruitment, the use of the terminology 'personalised medicine' may be misinterpreted by patients as a trial to provide them personalised care with respect to their best interest, rather than the actual purpose of finding scientific knowledge of a treatments efficacy.

1.4 Basket Trials

The work in this thesis focuses solely on basket clinical trials, considering various aspects of their design and analysis. Basket trials have been rapidly increasing in popularity amongst the rise of master protocol designs. A systematic review conducted by Park et al. (2019) found that the number of basket trials rose from just one prior to 2009, up to 49 by 2019. This rise has only continued since then. Of these 49 studies conducted by 2019, Park et al. (2019) found that 47 of them were in the oncology setting and 48 involved a drug investigation (the one exception was a trial for a vaccination against metastatic cancer). To add to this, 47 were exploratory phase I or phase II studies and only five contained a control group. A single control group in a basket trial can often be a difficult choice due to the range of disease types contained in a trial, with Park et al. (2020) stating that finding a common control arm across baskets may be infeasible. Therefore, to incorporate control groups in a basket trial, each basket would

have its own corresponding control arm with randomisation occurring within baskets, however, this is rarely seen implemented.

Basket trials have the practical advantage of containing several sub-studies under one protocol, which allows testing of treatments on rare conditions that would not typically warrant their own investigation due to their limited sample size and financial and time constraints that come along with RCTs. This is particularly important within cancer trials, as the number of different types of rare cancers is significant. Christyani et al. (2024) state that (as of publication in 2024) there are 230 distinct rare cancers. This further motivates the need for such basket trial designs in the oncology setting.

There are several prominent basket trials that have been implemented since 2009, one of which is the phase II VE-BASKET trial which tested Vemurafenib on cancers with the BRAF V600 mutation (Hyman et al., 2015). This genetic mutation is found across several tumour types including: non-small cell lung cancer, hairy-cell leukaemia, colorectal cancer and many more. The VE-BASKET trial ran from April 2012 to June 2014 and began with six cohorts of patients with various tumour types, whilst a seventh 'all other' cohort was also included, consisting of patients harbouring the BRAF V600 mutation but of a different disease type to the six established baskets. Later in the trial two new baskets were formed from the 'all other' sub-group based on sufficient enrolment for Erdhiem-Chester disease/Langerahan's cell histoctosis and anaplastic thyroid cancer, with the two groups warranting their own baskets. In contrast, three of the original six baskets were dropped due to a lack of enrolment, with patients then moved to the 'all other' group. Sample sizes across the baskets ranged from 7 to 27 patients, so even in the fastest recruiting basket, sample sizes were still limited. This comes down to the rarity of some cancer types, a common theme in basket trials but one that distinguishes itself from other study designs. The addition and closure of baskets in this VE-BASKET trial highlight the flexible nature of these studies. The VE-BASKET trial is referred back to and used as a motivating example throughout

this thesis, forming a basis for several simulation studies across Chapters 2-4.

Other examples of the implementation of basket trials is the MyPathway study (Hainsworth et al., 2018), which ran for 2014-2023. This trial consisted of multiple non-randomised basket trials under one overarching protocol, with each branch consisting of a basket trial based on a different genetic marker for metastatic solid tumours. One branch of this study looked at the combination of Vemurafenib on cancers with the BRAF V600 mutation, the same combination considered in the VE-BASKET trial. A further example of a phase II basket trial is the SUMMIT basket trial, which tested neratinib in patients with solid tumours harbouring a HER mutation (Hyman et al., 2018). Baskets included several tumour types including breast cancer, cervical cancer, non-small cell lung cancer and salivary gland cancer.

1.4.1 Hypotheses & Error Rates

As basket trials tend to be implemented in the early phases of drug development, the goal is often to test efficacy of the experimental treatment on each of the individual baskets on the trial. Throughout this thesis, binary responses are assumed and as such each patient either responds positively to the treatment or does not. That being said, the work throughout can be easily extended to continuous endpoints.

As responses are binary, they are modelled through a Binomial distribution with sample size, n_k, and unknown response rate, p_k, for basket k. Interest lies in the inference of these unknown response rates, thus the following family of hypotheses are tested:

$$H_0 : \ p_k \leq q_0 \quad vs. \quad H_1 : \ p_k > q_0, \quad \text{for } k = 1, \ldots, K,$$

where K is the total number of baskets on the trial and q_0 is the null response rate indicating a treatment is ineffective in a basket.

A Bayesian framework is often used in basket trials as it allows for incorporation of prior information which is combined with the observed response data, D, to produce a

posterior density (Muehlemann et al., 2023). From this posterior, probabilities can be computed, where a treatment is deemed effective in basket k if

$$\mathbb{P}(p_k > q_0|D) > \Delta.$$

The efficacy decision criteria, Δ, is calibrated to control some operating characteristic to a desirable level. This may be the basket-wise type I error rate (the probability of rejecting the null in a truly ineffective basket), the family-wise error rate (the probability of making at least one type I error across the K baskets) or the statistical power (the probability of rejecting the null in a truly effective basket). The calibration of such decision criteria is further explored in Chapter 3.

A common issue throughout all multi-arm studies, including basket trials, is the maintenance of both error rates and statistical power. When testing a family of hypotheses, the issue of multiplicity arises. Given stratified analysis of baskets, if a basket trial consists of K baskets with independent hypotheses and significance level α, then the FWER is given by:

$$\begin{aligned} \text{FWER} &= 1 - \mathbb{P}(\text{No true null hypotheses are rejected}) \\ &= 1 - (1-\alpha)^K. \end{aligned}$$

As such, the FWER rises to an unacceptable level as the number of baskets increases, this is highlighted in Figure 1.4.1. Several multiplicity correction methods exist to limit this inflation in FWER, these include Bonferroni correction, Dunnett's t-test and many others (Streiner, 2015).

The control of error rates to an appropriate level is vital for ensuring the validity of trial results. When baskets are analysed independently within a basket trial, independent hypothesis testing occurs and thus Howard et al. (2018) argue that no multiplicity corrections need to be implemented in order to maintain nominal error rates. However,

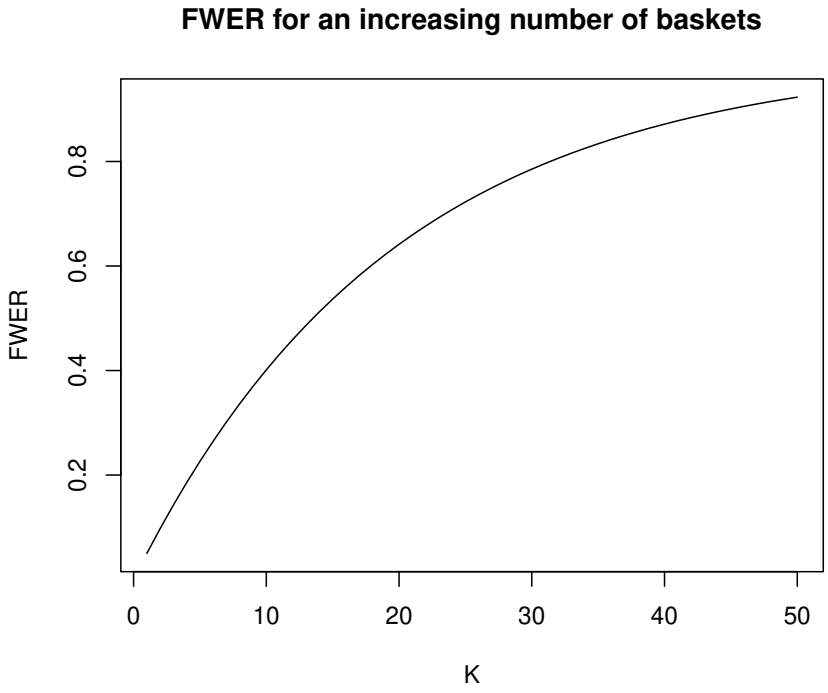


Figure 1.4.1: Inflation of the FWER as the number of baskets increase, with significance level $\alpha = 0.05$.

arguments are made against this throughout the literature. Cunanan et al. (2017) argue that even if baskets could be considered independent, they receive the same treatment and thus carry some form of correlation. Due to this relationship, they argue for the use of multiplicity corrections. Other design features such as the presence of a common control group, interim analysis, and early stopping for efficacy/futility need to be considered when making a decision on whether to correct or not. Throughout this thesis, no common control or interim analysis are conducted, therefore, the calibration of efficacy decision criteria are used for maintaining error rates.

Power is particularly difficult to maintain in basket trials, with the presence of small sample sizes causing uncertainty and increased variability in observed responses (Qin et al., 2019). This also results in potential biases and a lack of precision of treatment effect estimates. Other design characteristics can also have an impact on statistical power including the effect size, i.e. the difference between the null and target response

rate, and the significance level used. In the cases of rare disease it is often impossible to reach the level of recruitment required to maintain a sufficient level of power, thus creative Bayesian techniques are implemented in order to boost power (Kaizer et al., 2022).

1.4.2 Bayesian Information Borrowing

Extensive research has been conducted into the use of Bayesian methodology in basket trial designs. The goal of implementing such a framework is to improve power and precision of treatment effect estimates in the presence of small sample sizes. The basis of these methods revolve around the shared genetic component across all patients in a basket trial. This common genetic aberration leads to an 'exchangeability' assumption being made, which assumes a homogeneous response to treatment in all patients. Patients being 'exchangeable' means they can be moved between treatment baskets without changing the overall response rate estimates in each basket (Oakes, 2013) and as such, one can draw on information from one basket when making inference in another. This is known as Bayesian information borrowing.

As all baskets are assumed exchangeable, an extreme form of information borrowing is complete pooling of results. This combines results from all baskets and inference is made based on a single response rate, p, representing the probability of success for all baskets. This has the obvious disadvantage of losing any benefits gained from stratifying patients by disease group. In cases of homogeneity between all baskets, there is little issue with pooling, however in the presence of heterogeneity, as Neuenschwander et al. (2016) states, pooling bears the danger of overlooking baskets with interesting results. A basket trial testing the efficacy of Larotrectinib in patients with TRK Fusion–Positive Cancers (Drilon et al., 2018) took the approach of pooling response data from three phase I/II trials. This risked deeming the treatment effective in all three subgroups despite the possibility of the treatment being ineffective in one sub-study. This

could negatively impact patients, providing them treatment that may not be efficacious against their condition.

On the other end of the spectrum, although stratified analysis of baskets reduces the risk of heterogeneity between baskets, this form of analysis will lead to a lack of statistical power and precision in the presence of small sample sizes, as previously stated. Therefore an ideal analysis model falls between pooled and stratified analysis, adaptively borrowing between baskets based on the homogeneity of the response data.

There have been various Bayesian information borrowing methods proposed for the use in the basket trial setting, all of which involve an adaptive borrowing approach. Chapter 2 of this thesis provides an in depth comparison of several prominent approaches, exploring their characteristics through thorough simulation studies.

Although previously discussed in other areas of clinical trials, Berry et al. (2013) was the first to propose the use of Bayesian hierarchical models within phase II personalized medicine trials, of which basket trials are included. A Bayesian hierarchical model allows adaptive borrowing of information between sub-groups in a trial, with the amount of borrowing controlled by the degree of homogeneity of the observed data. A Bayesian hierarchical model (BHM) has the following form:

$$\begin{aligned} \theta_k &\sim \mathrm{N}(\mu, \sigma^2), \quad k = 1, \ldots, K, \\ \mu &\sim \mathrm{N}(m_\mu, \nu_\mu), \\ \sigma^2 &\sim g(\cdot), \end{aligned} \tag{1.4.1}$$

where θ_k could be the treatment effect in a controlled trial or the logit-transformed response rate. The common hyper-priors on the unknown mean and variance μ and σ^2 are shared amongst all baskets. Response rate estimates for each basket are shrunk towards their common mean, μ, with the degree of shrinkage controlled by the so called 'borrowing' parameter σ which reflects the heterogeneity between baskets. As σ^2 tends

to zero, borrowing moves towards complete pooling across baskets, whereas, as it tends to infinity borrowing becomes akin to stratified analysis. Although the model updates σ^2 based on the observed data, due to the small number of baskets often present, accurate estimation of between basket variation is challenging and, as stated by Berry et al. (2013), results in a high level of sensitivity to the choice of prior distribution, $g(\cdot)$. Gelman (2006) discussed prior options for σ^2, concluding that an Inverse-Gamma distribution as suggested by Berry et al. (2013), had poor behaviour around 0 and thus suggested a half-Cauchy prior on σ with a moderately large scale. In addition, subgroups with smaller sample sizes are expected to experience greater shrinkage to the mean, a particular issue in basket trials studying rare diseases.

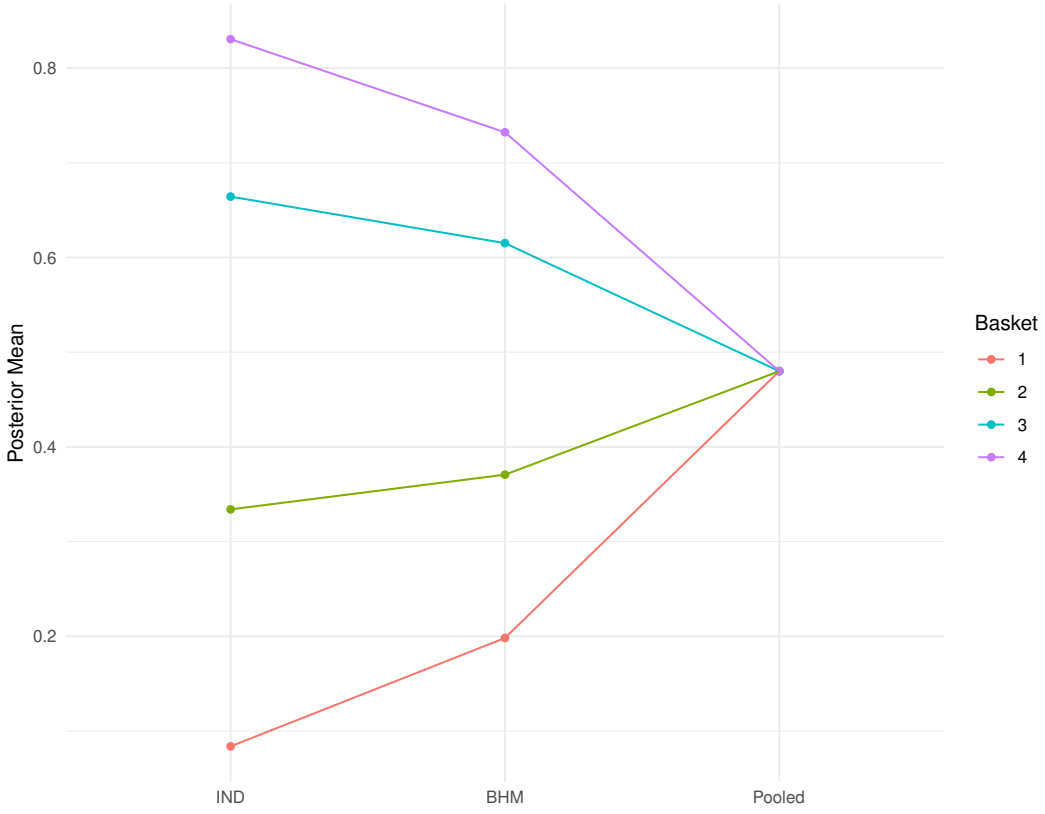


Figure 1.4.2: Mean posterior response rates of four basket with varying observed responses. Three models are fit to the data: an independent analysis (IND), a Bayesian hierarchical model (BHM) and a pooled analysis (Pooled).

The shrinkage towards the common mean can result in biased estimates and inflation

in error rates in cases of heterogeneity between baskets responses. This is highlighted in Figure 1.4.2, in which the BHM was applied to four baskets consisting of 12 patients each. The number of responses observed were 1, 4, 8 and 10 across the four baskets. Results under an independent analysis and complete pooling approach are also presented for comparison. The BHM shrinks the estimates towards the pooled estimate and away from those given by stratified analysis. For instance, basket 4 observed 10 responses out of 12 patients. Under an independent analysis this led to a posterior mean response rate for basket 4 of 0.83, whereas this was shrunk down to 0.73 under the BHM and 0.48 under a pooled analysis.

The issue of heterogeneity in information borrowing models is a well studied topic and various adaptations to the BHM have been made to better handle heterogeneous response data in basket trials. These methods include the calibrated Bayesian hierarchical model (CBMH, Chu and Yuan, 2018) and the exchangeability-nonexchangeability model (EXNEX, Neuenschwander et al., 2016). Alternative approaches include a Bayesian model averaging approach (BMA, Psioda et al., 2021) and Fujikawa's design (Fujikawa et al., 2020), all of which are discussed in depth in Chapters 2 and 4. Alternative options are the commensurate predictive prior approach (CPP, Zheng and Wason, 2022), a Bayesian hierarchical classification and information sharing approach (BaCIS, Chen and Lee, 2019), a multiple cohort expansion approach (MUCE, Lyu et al., 2023), robust Bayesian hypothesis testing (RoBoT, Zhou and Ji, 2020), Liu's two-path approach (Liu et al., 2017) and the Bayesian cluster hierarchical model (BCHM, Chen and Lee, 2020) to name a few.

1.4.3 Historic Information Borrowing

An alternative approach to improving power in trials with small sample size is to draw on information from historic or external data. As stated by van Rosmalen et al. (2018), a clinical trial is rarely performed in isolation, and in many cases, data from previous

studies are readily available. To utilise this historic information within a current trial, response data from the experimental treatment must be available, or alternatively, response information in a standard of care or control group from the patient population under investigation may also be used. It is historic control groups that are most readily available from previous studies. The use of historical control groups goes back decades. Kaizer et al. (2023b) discuss the use of historical controls in RCTs and state that the use of such controls was first introduced by Pocock (1976). However, official guidance by the FDA for the use of external controls was not published until February 2023 (U.S Food and Drug Administration, 2023a). Hall et al. (2021) discuss how in modern RCTs for rare diseases, the use of historical controls is an accepted practice.

Using historic data is particularly important when considering ethical implications of recruiting patients into control arms, for instance when a comparison to a placebo is desired but a more effective standard of care already exists (Marion and Althouse, 2023) or in cases with severe outcomes such as death, where no alternative treatment is available to the experimental treatment under investigation. They are also beneficial in cases of rare diseases, in which recruitment is difficult and it is impractical to enrol to a control arm within a current trial.

On the other hand, the use of historic control data does come with risks. Historic data is a potential source of bias should the patient population within the historic study differ inherently from the current patient population. This could be due to different evolution of diseases or illness, such as the change in variants throughout the COVID-19 pandemic (Marion and Althouse, 2023), or differing guidelines for treatment over time. Other sources of bias could be in the study design itself, which could include differences in inclusion/exclusion criteria, measurements of endpoints or treatment blinding approach.

Historic data is often used for designing a new study (Ghadessi et al., 2020), informing sample calculations, choosing endpoints and the null and target response rates

(Bennett, 2018). Under a Bayesian framework, the historic information is often incorporated into the prior distributions that are applied in a current trial. Several approaches have been used to define such prior distributions, with most involving down-weighting the information from the historical data, to account for heterogeneity between current and historic data sources (van Rosmalen et al., 2018). Several Bayesian approaches for borrowing information from historic controls are discussed in Chapter 4. Banbeta et al. (2019) and Bennett et al. (2021) also provide detailed comparisons of several historic information borrowing methods. The foundation of most of these methods is the power prior (PP), first introduced by Ibrahim and Chen (2000). This method raises the likelihood of the historical data to a fixed power, α. The power parameter, bound between 0 and 1, reflects the expected homogeneity between the historic and current data. Given that historic responses for basket k are denoted by y_{k^*}, the power prior has the form:

$$\pi(p_k|y_{k^*}, \alpha) \propto L(p_k|y_{k^*})^{\alpha}\pi_0(p_k),$$

where $\pi_0(p_k)$ is a vague prior on p_k before historic data is observed. There have been numerous extensions to this approach including, but not limited to: the modified power prior (MPP, Duan et al., 2006), the calibrated power prior (Pan et al., 2017), the meta-analytic-predictive prior (MAP, Zhang et al., 2021), a commensurate prior (Hobbs et al., 2011), the robust mixture prior (Schmidli et al., 2014) and a self-adapting mixture prior (SAM, Yang et al., 2023).

Su et al. (2022) discuss the implementation of such historic information borrowing in the precision medicine setting, particularly in the field of oncology trials. Su et al. (2022) compare such historic information borrowing to the approaches outlined in Section 2, borrowing information between current baskets on a trial. Baumann et al. (2023) applied the historic borrowing methods directly in the basket trial setting.

1.4.4 Adding Treatment Arms

The use of adaptive trial designs and master protocols allows modifications to the trial design while the study is still ongoing. Such modifications include interim analysis with futility and efficacy stopping, sample size adjustment or, as is investigated in this thesis, the addition of new treatment arms to an ongoing trial.

The addition of treatment arms was a critical component on the RECOVERY trial for treatments against the COVID-19 virus. In this trial, treatment arms were added or removed according to emerging evidence, whilst additional sub-studies were also added to provide more in depth information on secondary endpoints. Similarly, the I-SPY 2 trial, which is an ongoing adaptive platform trial investigating neoadjuvant therapy for patients with stage 2-3 breast cancer, allowed new treatments to be added to the trial at any interim time points (Wang and Yee, 2019).

The main benefit of adding treatment arms to an existing trial is the expedited trial process. Arms added under the current master protocol will benefit from the existing trial infrastructure and patient populations, which is a far more efficient approach then conducting separate trials for each new arm (Cohen et al., 2015). This will also prove beneficial for patients, speeding up how fast they receive potentially effective treatments. On the other hand, as we saw in Section 1.4.1, error rates only increase as the number of sub-groups grow, so any new groups added will bring another source of error and bias, making type I error rates more difficult to control. Korn and Freidlin (2017) discuss the practical difficulties of adding a treatment arm, stating that 'it is no minor undertaking' despite its efficiency. It brings along extra challenges such as regulatory complexity, the requirement of additional resources and the impact of staggered enrolment on the statistical analysis.

Although the addition of treatment arms has been a key feature of many platform trials, there is little mention of their use in the basket trial setting. This may be due to the rarity of the diseases typically under investigation in a basket clinical trial.

Assuming equal recruitment rates across baskets, any basket added at a later time point in the trial will suffer from even smaller sample sizes thus making inference challenging.

1.5 Thesis Outline

This thesis consists of three main content chapters. The overarching theme across chapters is the implementation of Bayesian information borrowing techniques in the basket trial setting. The first content chapter, Chapter 2 considers and compares existing Bayesian information borrowing models in basket trials, as well as, a proposal of an adaptation to one of these approaches. Chapter 3 investigates how to add new baskets to an ongoing basket trial with information borrowing applied. This chapter also explores the calibration of efficacy decision criteria, detailing the deficiencies in the traditional approach and outlining a novel robust procedure for calibrating these cut-off values. In Chapter 4, we investigate the use of historic information in basket trials and develop novel information borrowing models that incorporate borrowing from both historic and current baskets under one framework. Finally, in Chapter 5 we present a summary of our findings, make concluding remarks and outline several areas for future research. An overview of each chapter is provided below:

Chapter 2: A Comparison of Bayesian Information Borrowing Methods in Basket Trials and a Proposal of Modified EXNEX Method. In this chapter we review and compare the performance of several Bayesian borrowing methods, namely: the Bayesian hierarchical model (BHM), calibrated Bayesian hierarchical model (CBHM), exchangeability-nonexchangeability (EXNEX) model and a Bayesian model averaging (BMA) procedure. A generalisation of the CBHM is made to account for unequal sample sizes across baskets. We also propose a modification of the EXNEX model that allows for better control of a type I error. The proposed method uses a data-driven approach to account for the homogeneity of the response data, measured

through Hellinger distances. Through an extensive simulation study motivated by a real basket trial, for both equal and unequal sample sizes across baskets, we show that in the presence of a basket with a heterogeneous response, unlike the other methods discussed, this model can control type I error rates to a nominal level whilst yielding improved power.

Chapter 3: How to Add Baskets to an Ongoing Basket Trial with Information Borrowing. In this chapter, we explore approaches for adding baskets to an ongoing basket trial under Bayesian information borrowing and highlight when it is beneficial to add a basket compared to running a separate investigation for new baskets. We also propose a novel calibration approach for the decision criteria that is more robust to false decision making. Simulation studies are conducted to assess the performance of approaches which is monitored primarily through type I error control and precision of estimates. Results display a substantial improvement in power for a new basket when information borrowing is utilised, however, this comes with potential inflation of error rates which can be shown to be reduced under the proposed calibration procedure.

Chapter 4: Incorporating Historic Information to Further Improve Power When Conducting Bayesian Information Borrowing in Basket Trials. In this chapter we propose novel Bayesian methodology for incorporating historic or external data into a basket trial. It is well known that Bayesian information borrowing models can improve power and precision of estimates, however, an alternative approach is to incorporate any historical information available. This chapter considers models that amalgamate both forms of information borrowing, allowing borrowing between baskets in the ongoing trial whilst also drawing on response data from historical sources, with the aim to further improve treatment effect estimates. These models are data-driven, updating the degree of borrowing based on the level of homogeneity between information sources. A thorough simulation study is presented to draw comparisons between the

proposed approaches. We show that the incorporation of historic data under the novel approaches can lead to a substantial improvement in power of estimates when such data is homogeneous to the responses in the ongoing trial. Under some approaches, in cases of heterogeneity, this came alongside an inflation in type I error rate. However, the use of a power prior in the EXNEX model is shown to increase power and precision, whilst maintaining similar error rates to the standard EXNEX model in which no historic data is included.

Chapter 5: Conclusions and Further Work. This chapter provides an overview of the work presented in this thesis, summarising the key contributions and any limitations. In addition, this chapter highlights several areas of potential future research into the use of Bayesian methods in basket clinical trials.

Chapter 2

A Comparison of Bayesian Information Borrowing Methods in Basket Trials and a Proposal of Modified EXNEX Method

2.1 Introduction

Over the past decade there have been advancements in cancer genomics and refinement in diagnostic techniques, leading to the increased interest in the field of personalized medicine in which treatments are targeted to a specific genetic makeup (Lu et al., 2021). It would be infeasible to test these treatments on each of their targeted biomarkers in individual studies due to financial and time constraints. Master protocols have been proposed to tackle this problem. This term refers to trial designs that allow the testing of multiple treatments and/or multiple disease types in parallel under a single protocol (Bogin, 2020).

Basket trials are a form of master protocol that are usually implemented in phase II

of the drug development process within which a small number of patients are recruited to the study to determine efficacy of a treatment. Such a trial tests a single therapeutic treatment on several patient population sub-groups, each of which form a basket. Commonly, patients across all baskets share a genetic change/biomarker but each basket consists of patients with different diseases. One benefit of this trial design is its ability to test treatments which would traditionally not warrant their own investigation for their targeted patient population, due to their rarity and limited sample size.

As various groups in a basket trial share a common genetic aberration, a reasonable assumption can be made - known as the exchangeability assumption - that sub-groups may have a homogeneous response to the treatment (Jin et al., 2020). Specifically, the exchangeability assumption means that patients may be switched between exchangeable baskets without changing the overall value of the estimated basket treatment effects (Oakes, 2013). This exchangeability of patients across baskets implies that the response rates in all baskets can be viewed as random samples from the same model (Bernardo, 1996; Bernardo and Smith, 2009). There is some uncertainty surrounding the definition of nonexchangeability, in this thesis it is utilised to describe baskets between which no information is shared (usually due to heterogeneity in treatment effects). With this exchangeability assumption in mind, a concept known as 'information borrowing' can be used to draw on information regarding the response in one basket when estimating the response rate in others. This has the potential to increase power and precision of estimates, especially in the presence of small basket sample sizes. A desirable feature of such information borrowing methods is the ability to solely borrow between baskets with similar treatment effects, but not from those which are heterogeneous, as it may bias estimates and inflate the error rate resulting in a higher chance of a misleading conclusion. One would therefore like a method that has the ability to improve the power and precision of estimates while having control over error rates through only borrowing between homogeneous baskets.

Recently, numerous methods for information borrowing within the analysis of basket trials have been proposed. These methods either borrow information across all baskets such as the Bayesian hierarchical model (BHM, Berry et al., 2013) and the calibrated Bayesian hierarchical model (CBHM, Chu and Yuan, 2018), while others borrow between subsets of baskets, for example, the exchangeability nonexchangeability model (EXNEX, Neuenschwander et al., 2016) and a Bayesian model averaging approach (BMA, Psioda et al., 2021). This chapter provides a summary, alongside an extensive comparison of each method through simulation studies motivated by the VE-BASKET study, which consider both equal and unequal sample sizes across baskets. The consideration of unequal sample sizes is rare within the literature but an important aspect that needs to be considered when applying the models to clinical trial data.

We also propose an extension to the EXNEX model, which takes into account pairwise similarity between baskets' response rates through Hellinger distances in order to update the borrowing probability in the EXNEX model. The extension also involves excluding baskets with sufficiently heterogeneous responses to be treated as independent. In comparison to the EXNEX model, this method increases the sensitivity to the level of similarity between responses in order to borrow between homogeneous baskets with higher probabilities, whilst reducing the chance of borrowing from baskets with heterogeneous response rates in order to control the type I error rate to an appropriate level. We show that this proposed extension has the ability to increase power and precision of estimates compared to an independent/stratified analysis whilst controlling the type I error rate in some scenarios or performing similarly to the standard EXNEX model in others.

Although it may be clear that the performance of said information borrowing methods will depend on the homogeneity of the data, with methods that borrow information across all baskets outperforming those which borrow to a lesser extent in cases of homogeneity in response rates (and vice-versa under cases of heterogeneity), it is less clear

the impact this will have on certain operating characteristics such as error control. It is also a challenge to quantify the 'strength' of borrowing. The focus on this chapter is to monitor how certain metrics (primarily the type I error rate) are affected based on method used and homogeneity/heterogeneity of response data. This is explored through thorough simulation studies.

This chapter will be outlined as follows. In Section 2.1.1 we will introduce the setting of a motivating trial, the VE-BASKET study, that forms a basis for the comparison setting. In Section 2.2 we describe information borrowing models and propose the extension to the exchangalitiy-nonexchangibility model. In Section 2.3 we conduct a simulation study and then re-analyse the results of the VE-BASKET study using borrowing methods in Section 2.4.

2.1.1 Motivating Trial: VE-BASKET Study

This chapter is motivated by the VE-BASKET trial (Hyman et al., 2015) which explored the effect of Vemurafenib on multiple cancer types with the BRAFV600 mutation. From 2012 to 2014, 63 patients with the BRAFV600 mutation were enrolled and divided into baskets based on cancer types. The baskets included were non-small-cell lung cancer (NSCLC), Erdheim-Chester disease (ECD)/Langerhans'-cell histiocytosis (LCH), cholangiocarcinoma, colorectal cancer, anaplastic thyroid cancer and an 'all-other' group consisting of patients of different disease types with the BRAFV600 mutation. For the purpose of this work, baskets were only considered if they received the same treatment (Vemurafenib), with the same tumour criterion (solid tumour types) and thus the 'all-other' basket was excluded. The arms of the trial are summarised in Figure 2.1.1.

The primary endpoint of this study was the overall response rate (ORR) with a null response rate of 15% indicating inactivity. The target response rate was 45% while a response of 35% was considered low but still indicative of a response. For a stratified

analysis of baskets, the planned sample size, obtained through a Simon's two-stage design (Simon, 1989), was 13 per basket based on 80% power and 10% type I error rate. However, different sample sizes were realised with the Thyroid cancer basket, for example, consisting of just 7 patients. This limited sample size causes issues when drawing inference from trial results as estimation of treatment effects will lack precision and thus any conclusions made regarding the effect of Vemurafenib on thyroid cancer may be questionable. However, due to baskets sharing a common genetic aberration one can utilise information borrowing techniques.

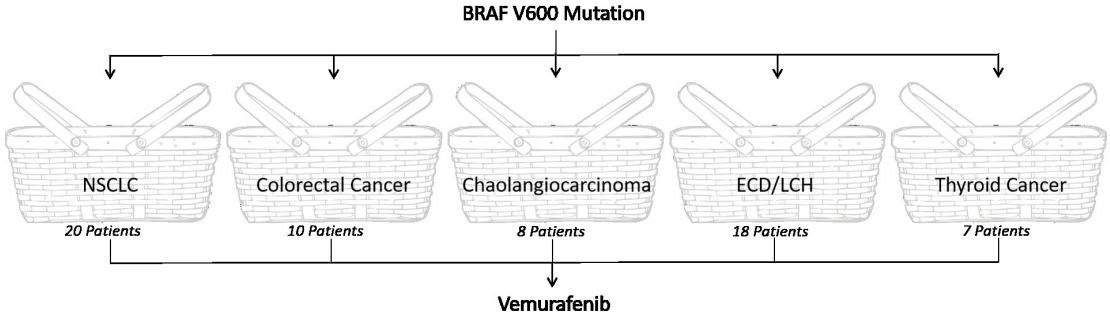


Figure 2.1.1: VE-BASKET trial design: the 5 baskets included in the study alongside their observed sample sizes.

2.2 Methods

2.2.1 Setting

Consider a basket trial consisting of K baskets. This chapter focuses on a single treatment arm setting and a primary binary endpoint, in which a patient either responds positively to a treatment or does not. Denote the responses in basket k ($k = 1, \ldots, K$) by Y_k, which follows a binomial distribution, $Y_k \sim \text{Bin}(n_k, p_k)$, with n_k and p_k indicating the sample size and response rate in basket k respectively. Interest lies in estimating the unknown response rate, p_k. Denote q_0 as the null response rate which indicates inactivity and q_1 as the target response rate. The objective is to test the family of

hypotheses:

$$H_0 \ : \ p_k \leq q_0 \qquad vs. \qquad H_1 \ : \ p_k > q_0, \qquad k = 1, \ldots, K.$$

To test these hypotheses a Bayesian framework is used. Having observed data D, at the conclusion of the trial the treatment is deemed effective in basket k if $\mathbb{P}(p_k > q_0|D) > \Delta_\alpha$.

The decision cut-off, Δ_α, is typically calibrated under a null scenario in which the treatment effect is homogeneous and ineffective across baskets, to control error rates at a nominal level, α. This chapter utilises calibration in order to control a basket specific type I error at the nominal level under a null scenario, however, as an alternative approach Psioda et al. (2021) instead calibrated to control the family-wise error rate across all baskets in the trial. Despite this calibration, methods that borrow information from heterogeneous baskets are expected to have error rates greater than α. Borrowing causes a shift in the posterior density of p_k towards a common mean and thus, when borrowing from a basket with a larger heterogeneous response, the point estimate obtained tends to increase, as does the probability $\mathbb{P}(p_k > q_0|D)$, so more baskets are erroneously deemed sensitive to treatment. When no borrowing occurs this shift is not present as the level of heterogeneity is irrelevant, so the same inflation is not expected.

2.2.2 Independent Model

Independent analysis is an approach that does not borrow information between baskets and instead conducts stratified analysis for each. As such, for each basket, only data observed from its set of patients is considered when estimating its treatment effect. For

each basket k in $1, \ldots, K$

$$\begin{aligned} Y_k &\sim \text{Binomial}(n_k, p_k), \\ \theta_k &= \log\left(\frac{p_k}{1-p_k}\right), \\ \theta_k &\sim \text{N}(\text{logit}(q_{0k}), \nu_k), \end{aligned} \tag{2.2.1}$$

where q_{0k} denotes the null response rate in basket k. The logit transformation of the response rates is taken to avoid boundary issues when p_k is close to 0 or 1 and to align with the borrowing models to allow for a fair comparison. A slightly informative normal prior is placed on this transformed parameter, with mean based on the null response rate but with a large variance, ν_k. This method controls the type I error rate as the response rates do not depend on the level of heterogeneity across baskets, but estimates lack statistical power and suffer lower precision when a basket has a small sample size (Cunanan et al., 2018).

2.2.3 Bayesian Hierarchical Model

The Bayesian hierarchical model (BHM), proposed by Berry et al. (2013), utilises the full exchangeability assumption as all baskets share a common genetic change. With this assumption in mind, each basket's response to a treatment can be expected to be homogeneous and thus information can be shared between all baskets in the trial. The BHM is specified such that the log-odds of the response rate for each basket follows a normal distribution, centred around a common mean μ with variance σ^2. Hyper-priors

are placed on the parameters μ and σ^2.

$$\begin{aligned} Y_k &\sim \text{Binomial}(n_k, p_k), \qquad k = 1, \ldots, K \\ \theta_k &= \log\left(\frac{p_k}{1-p_k}\right) \sim N(\mu, \sigma^2), \\ \mu &\sim N(\text{logit}(q_0), \nu_\mu), \quad \sigma \sim g(\cdot). \end{aligned} \tag{2.2.2}$$

The hyper-prior on μ is suggested to be slightly informative (Berry et al., 2013) based on the average null response rate across the baskets, with a large variance. The choice of hyper-prior on σ, $g(\cdot)$, is widely debated with Inverse-Gamma, Half-Normal or Half-Cauchy densities commonly used. An Inverse-Gamma prior on σ^2 was utilised in the original paper, however, as stated by Gelman (2006), this has poor behaviour when σ^2 is close to 0 and thus a half-Cauchy prior on σ with a moderately large scale was suggested instead.

Under the BHM, borrowing occurs between all baskets and as a result, the estimates of the response rates for each basket are shrunk towards the common mean with the degree of shrinkage controlled by the so called shrinkage/borrowing parameter, σ^2. When σ^2 tends to 0, borrowing moves towards the complete pooling approach in which the results of all baskets are combined and inference is made based on a single response rate. At the other extreme, when σ^2 tends to infinity, inference is akin to an independent analysis. This pull towards the common mean can result in a basket's treatment effect estimate being pulled away from the true value, particularly in the presence of a heterogeneous basket.

2.2.4 Calibrated Bayesian Hierarchical Model

The calibrated Bayesian hierarchical model (CBHM), proposed by Chu and Yuan (2018) is an extension of the BHM and as such also makes the full exchangeability assumption. The CBHM has the same form as Model (2.2.2), but rather than placing a prior on

σ directly, it is defined as a function of a measure of homogeneity across baskets: $\sigma^2 = \exp\{a + b\log(T)\}$, where T is the chi-squared test statistic for homogeneity:

$$T = \sum_{k=1}^{K} \frac{(O_{0k} - E_{0k})^2}{E_{0k}} + \sum_{k=1}^{K} \frac{(O_{1k} - E_{1k})^2}{E_{1k}}, \tag{2.2.3}$$

where O_{0k} and O_{1k} are the observed failures and responses in basket k respectively, while E_{0k} and E_{1k} are the expected failures and responses in basket k.

The parameters a and b are tuned to calibrate the function to ensure strong borrowing through hierarchical modelling when all baskets have a homogeneous response and treat baskets as independent otherwise. The calibration procedure is outlined by Chu and Yuan (2018) as follows:

1. Generate W simulated data sets in which the treatment is effective in all baskets' with response rate q_1, for each computing T as in (2.2.3). Let H_B be the median of these T values.

2. Simulate the case in which the treatment effect is heterogeneous across baskets. To do so, let $q(j) = (q_1, \ldots, q_1, q_0, \ldots, q_0)$ be the scenario in which the treatment is effective in the first j baskets but not effective in baskets $j+1$ to K. For each value of $j \in \{1, \ldots, K-1\}$ generate W simulations of data, calculating the test statistic T for each. Denote $H_{\bar{B}j}$ as the median value of T for each value of j. Finally, define $H_{\bar{B}} = \min_j(H_{\bar{B}j})$.

3. A small value of σ_B^2 is chosen to reflect strong information borrowing when the responses are homogeneous across the baskets and a large value $\sigma_{\bar{B}}^2$ is also selected to reflect weak information borrowing when the responses are heterogeneous across all/some baskets. As suggested by Chu and Yuan (2018), these values could be $\sigma_B^2 = 1$ and $\sigma_{\bar{B}}^2 = 80$. These values are then used to solve for a and b for $\sigma_B^2 = g(H_B)$ and $\sigma_{\bar{B}}^2 = g(H_{\bar{B}})$, where $\sigma^2 = g(T) = \exp\{a + b\log(T)\}$. This results

in:

$$a = \log(\sigma_B^2) - \frac{\log(\sigma_{\bar{B}}^2) - \log(\sigma_B^2)}{\log(H_{\bar{B}}) - \log(H_B)} \log(H_B), \qquad b = \frac{\log(\sigma_{\bar{B}}^2) - \log(\sigma_B^2)}{\log(H_{\bar{B}}) - \log(H_B)}.$$

A benefit of such a tuning procedure is the increased certainty in estimates produced by the CBHM in comparison to the BHM in the case where all baskets are homogeneous. However, with a and b tuned in this way, the method takes on a 'strong' definition of heterogeneity such that if the response rate in one basket is heterogeneous, then all baskets are deemed heterogeneous, and as a result no borrowing occurs. The 'strong' definition of heterogeneity can be relaxed through a less stringent tuning procedure but this comes at the cost of the error control.

Note that Chu and Yuan (2018) explored the sensitivity of results to the choice of σ_B^2 and $\sigma_{\bar{B}}^2$ to reflect weak and strong borrowing in step 3 of the calibration procedure. They found that as long as the choices were reasonably small and large respectively, the choice of values had little impact on performance.

The original calibration procedure for the CBHM, proposed by Chu and Yuan (2018) was based on equal sample sizes for each basket. In practice it is unlikely that all baskets will recruit exactly the same number of patients, so the calibration outlined above may not be adequate. When the sample sizes differ, step 2 in the calibration does not cover all possibilities of heterogeneity as the ordering of response rates matter. We propose altering this step for unequal sample sizes to consider all permutations of q_1 and q_0 in which at least one basket has response rate q_0 and at least one has response rate q_1.

2.2.5 Exchangeability-Nonexchangeability Model

The full exchangeability assumption is often violated in the presence of heterogeneous baskets. The exchangeability-nonexchangeability (EXNEX) model, proposed by Neuenschwander et al. (2016) incorporates a nonexchangeability component to the standard

Bayesian hierarchical model, within which no borrowing occurs. The model then has two components:

1. EX (exchangeable component): with prior probability π_k, basket k is exchangeable and a Bayesian hierarchical model as in model (2.2.2) is applied. Information borrowing is therefore conducted between all baskets assigned to the exchangeable component.

2. NEX (nonexchangeable component): with prior probability $1 - \pi_k$, θ_k is nonexchangeable with any other basket, and as a result, basket k is treated independently.

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && M_{1k} \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)} \\
&\theta_k = \log\Big(\frac{p_k}{1-p_k}\Big), && \mu \sim \text{N}(\text{logit}(q_0), \nu_\mu), \\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && \sigma \sim g(\cdot), \\
&\delta_k \sim \text{Bernoulli}(\pi_k), && M_{2k} \sim N(m_k, \nu_k). \quad \text{(NEX)}
\end{aligned} \tag{2.2.4}
$$

As information is borrowed only between baskets assigned to the EX component but not from those in the NEX component, this model provides more flexibility compared to the previous methods as information can be borrowed between just some of the baskets and not all of them.

Careful consideration is needed in this model when it comes to the selection of π_k values. It is uncommon to have strong prior information on the probability of exchangeability, so it is suggested to fix these prior to the trial at $\pi_k = 0.5$ for all baskets. This prior probability is updated to some degree based on the homogeneity of the data but is not sensitive enough to the heterogeneity/homogeneity of responses and thus it is anticipated that the probability of borrowing from a heterogeneous basket will be too high, which in turn will inflate the type I error rate. Ideally the prior probability of assigning homogeneous baskets to the exchangeability component should increase,

while those for heterogeneous baskets decreases as opposed to fixing these probabilities at 0.5 each.

Note that a Dirichlet prior could be placed on π_k, however, as stated by Neuenschwander et al. (2016) this does not have a substantial effect on inference in comparison to fixing the weights a priori. The EXNEX model can also be easily extended to have more than one exchangeability component, allowing us to borrow between different subsets of baskets.

2.2.6 Proposed Modified EXNEX Model

In the original EXNEX model, the prior probability values, π_k, do not depend on the similarity of the data. We propose a modification to the EXNEX model, denoted mEXNEX$_c$, which sets these π_k values to account for the homogeneity of the response in basket k compared to that in all other baskets. A similar concept of updating prior weights based on homogeneity of responses was proposed by Zheng and Hampson (2020) but in the dose-finding setting. The purpose of this is to increase the sensitivity to the heterogeneity of response data compared to the EXNEX model.

The Hellinger distance is an ideal metric that quantifies the similarity between two probability distributions parameterised by probability density functions. In the mEXNEX$_c$ model it is used to compare the distance in responses between baskets. The Hellinger distance gives values on the $[0, 1]$ range, equating to 0 when densities are identical and increasing values as the distance between the densities becomes greater and as such, they can be easily translated into probability values.

The mEXNEX$_c$ model is a two-step procedure, the first step removes baskets with a clearly heterogeneous response rate. A pre-specified cut-off value, c, is chosen to indicate that a basket is sufficiently heterogeneous to exclude from borrowing and treat as independent. Denote $\hat{p}_k = Y_k/n_k$. If the minimum pairwise difference in response

rate between basket k and all other baskets is greater than c,

$$\min_{k'}\{\ |\hat{p}_k - \hat{p}_{k'}|\ \} > c, \quad k \neq k',$$

then basket k is treated as independent and its mixture weight, π_k, in the EXNEX model is set to 0.

In the second step, denote S as the set of all baskets not excluded for heterogeneity. For all baskets in S, produce posterior densities for p_k by fitting a beta-binomial model with prior $p_k \sim \text{Beta}(1,1)$, which has form $p_k|Y_k \sim \text{Beta}(a_k, b_k)$ where $a_k = Y_k + 1$ and $b_k = n_k - Y_k + 1$. The Hellinger distance between posteriors of basket k and k' is computed as

$$h_{k,k'} = \sqrt{1 - \frac{B\left(\frac{a_k+a_{k'}}{2}, \frac{b_k+b_{k'}}{2}\right)}{\sqrt{B(a_k,b_k)B(a_{k'},b_{k'})}}}, \quad (k,k' \in S) \tag{2.2.5}$$

where $B(\cdot,\cdot)$ is the Beta function. The probability, π_k, is then calculated as

$$\pi_k = \sum_{k'} \frac{1 - h_{k,k'}}{|S| - 1} \ \text{ for } \ k,k' \in S, \ \ k \neq k'.$$

Once obtained, these π_k values are used as the prior borrowing probabilities in model (2.2.4). For the mEXNEX$_c$ model, a slight alteration is made to model (2.2.4), in that, a prior is placed on σ^2 as opposed to σ in order to have less mass concentrated around 0.

This method is expected to reduce the probability of heterogeneous baskets being assigned to the EX component as a heterogeneous basket will have larger Hellinger distances and thus lower π_k values. As such, the mEXNEX$_c$ model is expected to possess better error control than the standard EXNEX model that assigns fixed π_k values irrespective of the homogeneity of responses.

The specification of the cut-off c to define a basket as sufficiently heterogeneous

to remove requires careful consideration. When defining c prior to the trial, the clinician must weigh up the trade-off between achieving higher power of estimates while maintaining an adequate error rate. A larger c value will result in higher power at the cost of inflation of error rates, whilst lower, more conservative values control error rates but provide a smaller increase in power. A cut-off is chosen such that this trade-off is considered acceptable.

A proposed method for this specification is through a pre-trial simulation study in which the null and target response rate and planned samples sizes are used to compute operating characteristics for different values of c, with Δ_α re-calibrated for each. The planned sample sizes are obtained as in the trial protocol, using a Simon two-stage design based on stratified analysis on each basket for a targeted type I error rate and power. Generally, consider cut-off values of $c = i/\max n_k$ for $i = 0, 1, 2, \ldots, n_k$ and $k = 1, \ldots, K$ to reflect all possible differences in point estimates (ranging from 0 responses out of a sample n_k up to all n_k patients responding). Also, it is important to include data scenarios that cover all combination of insensitive and sensitive baskets.

A utility function is provided in order to guide the selection of c based on the pre-trial simulation results:

$$c = \arg\max_c \{\ x\text{Power}_c + (1-x)(1-\text{Error}_c)\ \}, \quad x \in [0,1], \tag{2.2.6}$$

where Power_c and Error_c are the mean power and type I error rate for cut-off c across all considered scenarios. The value of x is chosen subjectively by a clinician to reflect the importance of type I error-rate control or power improvement in the specific basket trial application. A larger value of x would place more emphasis on power improvement and less on type I error control, whilst smaller values of x would place more emphasis on error control than power improvement. Due to the trade-off of power and error rate, if x is chosen to be too large, the type I error will likely inflate over the nominal level in order to maximise the power. Similarly, if x is chosen to be too small, the borrowing

would be overly conservative and thus power may not reach the targeted value.

2.2.7 Bayesian Model Averaging

Psioda et al. (2021) proposed a Bayesian model averaging (BMA) approach that allows for both exchangeability and nonexchangeability, but in place of applying a single model to the data, an average is taken over all of the considered models. To do so one averages over the posterior distribution under each model, weighted by their posterior model probability (Hoeting et al., 1999).

Consider the case where only a single exchangeability component is allowed. Define $\mathcal{M}_j$ as model j representing a permutation of basket allocation to the EX group or NEX group. Rather than applying a hierarchical model to borrow between baskets in the EX group, results are pooled and baskets have one shared response rate $p_{S_{i,j}}$, where $S_{i,j}$ is a subset i of the baskets' given model $\mathcal{M}_j$. Therefore, $p_k = p_{S_{i,j}}$ when $k \in S_{i,j}$.

A weakly informative Beta prior is placed on the response rates, while a prior on each model, $f(\mathcal{M}_j)$, is also required. The posteriors $f(p_k|\mathcal{M}_j)$ and $f(\mathcal{M}_j|D)$ are computed after observing response data D and are used to implement a BMA procedure to obtain the efficacy decision for basket k at the conclusion of a trial by computing $\mathbb{P}(p_k > x|D) = \sum_j \mathbb{P}(p_k > x|\mathcal{M}_j, D) f(\mathcal{M}_j|D)$.

This method is potentially advantageous as it accounts for all possible borrowing subsets in place of applying a single model. This allows for uncertainty in the model selection, as the specification of an incorrect model may lead to misleading inference. Also, as a result of pooling within exchangeability groups, closed-form solutions of posteriors can be found. This is computationally appealing as it can be implemented quickly even for a large number of baskets.

2.3 Simulation Study

In order to assess the performance of the described methods in terms of estimation, type I error and power, a simulation study was conducted. Motivated by the VE-BASKET trial, the conducted simulation study consists of 5 baskets. Two settings are considered:

(i) Sample size in each basket being equal to the planned sample size of 13 patients,

(ii) Sample sizes in the baskets being the realised sample sizes in the trial (i.e. 20, 10, 8, 18 and 7).

Set $q_0 = 0.15$ and $q_1 = 0.45$ as the null and target response rates respectively. A basket is deemed sensitive to a treatment at the conclusion of a trial, having observed data D, if $\mathbb{P}(p_k \geq 0.15|D) > \Delta_\alpha$, where Δ_α is calibrated to obtain a type I error rate of $\alpha = 10\%$ under the null scenario. Note that Δ_α is calibrated for each method separately and follows the same procedure for both the proposed and existing methods - for the mEXNEX$_c$ model, c is selected through calibration but is then taken as fixed when calibrating Δ_α. This is done based on the planned sample size $n_k = 13$ for all baskets k and the null response rate $q_0 = 0.15$. The calibrated Δ_α values for each method are given in Table A.1.1 in the Supporting Information.

Several scenarios with varying numbers of baskets sensitive to treatment are considered and displayed in Table 2.3.1. Scenario 1 is the null case in which all baskets are insensitive. Scenarios 2-5 cover different combinations of insensitive and sensitive treatment baskets while scenario 6 is the case where all baskets are homogeneous and sensitive. This will highlight the benefits, if any, the borrowing methods provide in terms of power improvement. Scenarios 7-10 consist of cases where some baskets have a marginally effective response rate at 35%. For the realised sample size case, a further 6 data scenarios are considered to account for the fact that ordering of response rate now matters.

For each method and scenario the following operating characteristics are computed:

Table 2.3.1: True response rate data scenarios for comparison of information borrowing models. For the planned sample size simulation, scenarios 1-10 are considered, whereas, for the realised sample size simulation all scenarios 1-16 are considered.

	p_1	p_2	p_3	p_4	p_5
Scenario 1	0.15	0.15	0.15	0.15	0.15
Scenario 2	0.45	0.15	0.15	0.15	0.15
Scenario 3	0.45	0.45	0.15	0.15	0.15
Scenario 4	0.45	0.45	0.45	0.15	0.15
Scenario 5	0.45	0.45	0.45	0.45	0.15
Scenario 6	0.45	0.45	0.45	0.45	0.45
Scenario 7	0.35	0.15	0.15	0.15	0.15
Scenario 8	0.35	0.35	0.35	0.15	0.15
Scenario 9	0.45	0.35	0.35	0.15	0.15
Scenario 10	0.45	0.45	0.35	0.35	0.15
Scenario 11	0.15	0.15	0.15	0.15	0.45
Scenario 12	0.15	0.15	0.45	0.15	0.45
Scenario 13	0.15	0.45	0.45	0.15	0.45
Scenario 14	0.15	0.45	0.45	0.45	0.45
Scenario 15	0.45	0.15	0.15	0.15	0.45
Scenario 16	0.45	0.15	0.45	0.15	0.45

- % Reject: the percentage of simulated data sets in which the null hypothesis is rejected. If the null is true then this value is the type I error rate, else it is the power.

- % All Correct: the percentage of simulated data sets in which the correct conclusions are made across all baskets.

- FWER (family-wise error rate): the percentage of simulated data sets in which at least one null basket is deemed sensitive to treatment.

- Mean point estimate of the response rate in each basket and the standard deviation of said estimate across the simulations.

The results presented focus on the first three of these, with results for the mean point estimates provided in Section A.3 of the Supporting Information.

For the following analysis, prior and parameter choices for each model are sum-

Table 2.3.2: Model prior and parameter specification for the simulation study to compare information borrowing methods.

Model	**Prior & Parameter Specification**
Independent	$\theta_k \sim \text{N}(\text{logit}(0.15), 10^2)$
BHM	$\mu \sim \text{N}(\text{logit}(0.15), 10^2)$, $\sigma \sim \text{Half-Cauchy}(0, 25)$
CBHM	$\mu \sim \text{N}(\text{logit}(0.15), 10^2)$, $\sigma^2 = \exp\{-7.25 + 5.86 \log(T)\}$
EXNEX	$\mu \sim \text{N}(\text{logit}(0.15), 10^2)$, $\sigma \sim \text{Half-Normal}(0, 1)$, $M_{2k} \sim \text{N}(-0.62, 4.4^2)$, $\delta_k \sim \text{Bernoulli}(0.5)$
mEXNEX$_c$	$\mu \sim \text{N}(\text{logit}(0.15), 10^2)$, $\sigma^2 \sim \text{Half-Normal}(0, 1)$, $M_{2k} \sim \text{N}(-0.62, 4.4^2)$, $\delta_k \sim \text{Bernoulli}(\pi_k)$
BMA	$P_{S_j} \vert \mathcal{M}_j \sim \text{Beta}(0.45, 0.55)$, $f(\mathcal{M}_j) \sim P_j^2$

marised in Table 2.3.2 with full model specification provided in Appendix 2.6.1. Priors on μ are centred around the null response rate of 0.15 with a large variance. Priors on σ^2 are chosen to be consistent with those used in the literature. The EXNEX model has prior borrowing probabilities fixed at 0.5. The prior parameters for the mEXNEXc model are kept the same as the standard EXNEX model to allow for fairer comparison. These parameters are selected by the recommendation of Neuenschwander et al. (2016) with the prior for the NEX component in both the EXNEX and the mEXNEX$_c$ model centred around a plausible guess of p_k of 0.35. In the BMA only a single EX group is implemented and the priors are consistent with those suggested by Psioda et al. (2021). The prior placed on model $\mathcal{M}_j$ is $f(\mathcal{M}_j) \propto P_j^2$, where P_j^2 is the total number of distinct response rates in model $\mathcal{M}_j$ (the EX component has a shared response rate and all baskets in the NEX component have a distinct response rate, thus P_j is the number of baskets in the NEX component plus 1).

The specification of the cut-off value, c, in the mEXNEX$_c$ model is chosen through a pre-trial simulation as outlined in Section 2.2.6. Cut-off values of $c = i/13$ were considered where $i = 0, 1, 2, 3, 4$. For each value of c, 10,000 simulated data sets were

used to compute the type I error rate and power across the 6 scenarios in Table 2.3.1, with the results shown in Figure 2.3.1.

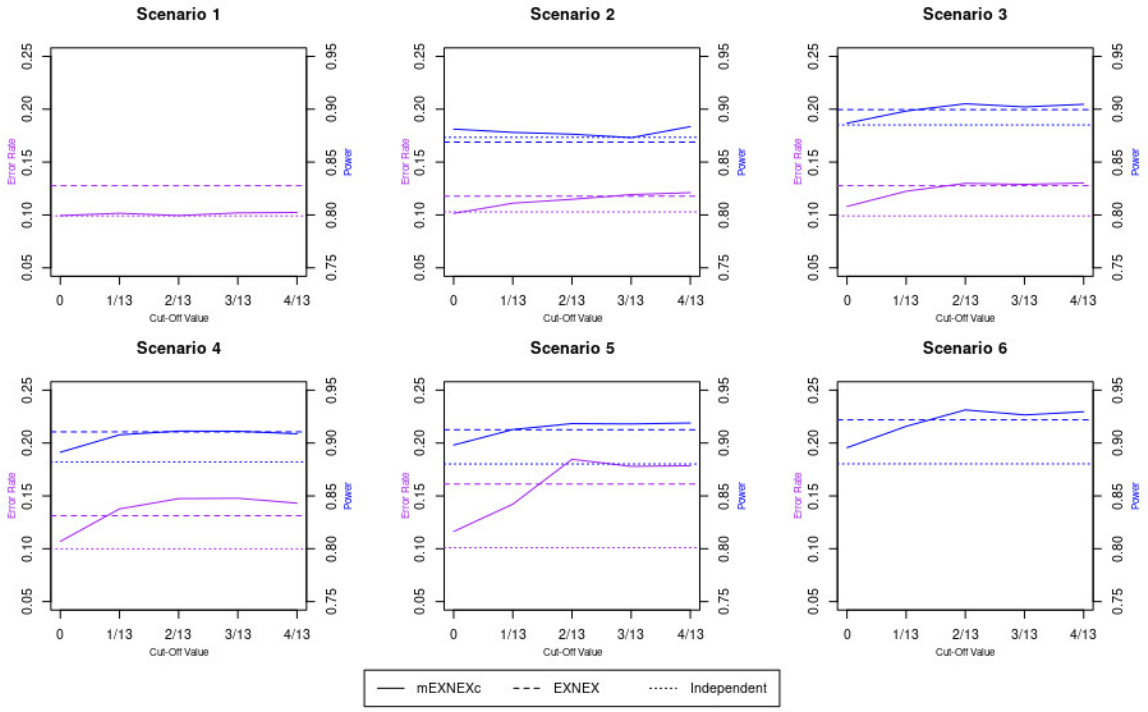


Figure 2.3.1: Pre-trial simulation results for type I error rate and power across the data scenarios under the mEXNEX$_c$ model for different cut-off values, c.

Within Equation (2.2.6) two cases were considered: when $x = 0.4$ a higher emphasis is placed on error control over power improvement resulting in the choice $c = 0$. This is a more conservative value as it only allows for borrowing when a basket has an identical response rate to at least one other basket. However, despite this conservative nature, from Figure 2.3.1, we observe that this specification shows control of the type I error rate close to the nominal 10% level under scenarios 1-4, whilst improving power under scenarios 2, 4, 5 and 6 in comparison to an independent model. Denote this model as mEXNEX$_0$. The second choice is $x = 0.6$ which puts greater weight on power improvement whilst relaxing the degree of error control, resulting in the choice $c = 1/13$. Denote this model as mEXNEX$_{1/13}$. Both choices of x, 0.4 and 0.6, are selected to be close to 0.5 in order to avoid imbalance in power improvement and type I error inflation.

A total of 10,000 simulations were run using the 'rjags' package v 4.12, (Plummer, 2023) within RStudio v 1.1.453 (R Core Team, 2020) for each of the 6 data scenarios in Table 2.3.1.

2.3.1 Simulation Results: Planned Sample Sizes

The results for power and type I error rate under the planned sample size are presented in Figure 2.3.2 which shows the percentage of simulated data sets in which the null hypothesis was rejected for each method and scenario. Full results are also provided in Table A.1.2 and A.1.3 in the Supporting Information.

The rejection percentages are calibrated under scenario 1 to achieve a 10% type I error rate for each method separately and hence all rejections are approximately 10%. However, in the presence of a single heterogeneous effective basket, i.e. scenario 2, the mEXNEX$_0$ model gives the best performance with error control at 10% whilst achieving the greatest power (88%). The CBHM also controls the type I error rate but only achieves 81.1% power due to the level of heterogeneity and the nature of the calibration procedure. The BHM, BMA and EXNEX model all have raised error rates at approximately 16.9%, 13.2% and 11.8% respectively. The mEXNEX$_{1/13}$ model gives power that is increased by 0.8 compared to the EXNEX model with a slightly lower type I error rate of 11.5%.

Across scenarios 3, 4 and 5 there is a mix of sensitive and insensitive treatment baskets. These scenarios show the benefits in terms of power gain through information borrowing techniques compared to an independent analysis. Again in these cases the CBHM lacks power due to the heterogeneity of the data, giving consistently lower power than an independent analysis, whilst the BHM and BMA procedure give hugely inflated error rates. The BHM gives type I error rates ranging from 21.6% to 42.1% across these three scenarios.

The mEXNEX$_{1/13}$ model leads to similar results to the standard EXNEX model

Figure 2.3.2: Percentage of rejections of the null hypothesis for each information borrowing method and data scenario based on a planned sample size of 13 patients per basket.

in these scenarios due to its inability to detect clusters of responses, which leads to increased probability values for EX assignment for all baskets. In scenario 4 this results in a greater type I error rate for the mEXNEX$_{1/13}$ compared to the EXNEX model (15% vs. 13.1%), however, under scenario 5 the mEXNEX$_{1/13}$ model gives a 1.3 decrease in the type I error rate. Under a more conservative cut-off, the mEXNEX$_0$ model keeps the type I error rate at an acceptable level with the worst case occurring under scenario 5 in which the error rate is just 11.2%, which is much lower than the 16.1% error rate of the EXNEX model. This is all whilst also increasing power over an independent model by 1.4%.

In scenario 6, when all baskets are sensitive to treatment, the BHM followed by the BMA procedure give the greatest power at the cost of inflated error rates across the other scenarios. The mEXNEX$_{1/13}$ model has similar power to the standard EXNEX model with mean power 91.9% compared to 92.2%, whereas the mEXNEX$_0$ model, has lower average power at 89.8% but still an improvement over the independent model at 88.0%.

Now consider the cases where some baskets are marginally effective with a true response rate of 35%. In particular one can draw comparisons between scenario 2 and 7 as in both cases just a single basket is heterogeneous and effective to some degree. Under both scenarios the same patterns of results are observed, but due to the lower true response rate under scenario 7, the difference in power and error rates between methods has been amplified. As expected, the error rates tend to be lower under scenario 7 compared to 2, as the pull upwards towards the heterogeneous basket will be less extreme as it has a true response rate closer to that under the null. In this case, both mEXNEX$_c$ models give the joint highest power at 68.3%, with the mEXNEX$_0$ model again controlling error rates at the 10% level, while only minimal inflation is observed under the mEXNEX$_{1/13}$ model at 11.1% (a value very similar to that of the standard EXNEX model). The BHM, CBHM and BMA approach all give lower power

than an independent analysis with clearly inflated error rates in the BHM and BMA cases. Similar connections can be made between scenarios 4 and 8, with the same conclusions drawn from each.

Under both scenarios 9 and 10, baskets have a combination of effective, marginally effective and ineffective response rates. Predictably, the BHM and BMA approach give the greatest power but with this have inflated error rates, just as in scenarios 4 and 5. All of these scenarios 7-10 demonstrate the ability of the mEXNEX$_c$ model to control error rates when $c = 0$ whilst improving power over an independent analysis anywhere from 1.8-3% for effective baskets and 3.3-4.5% for marginally effective baskets.

Looking now at the percentage of data sets in which the correct inference was made across all baskets, alongside the family-wise error rates (where Δ_α was now calibrated under scenario 1 to achieve 25% FWER - full results provided in the Supporting Information (Section A.1)). All methods gave similar values for correct inference under the null scenario. However, under both scenarios 2 and 7, the independent model produced the greatest values, closely followed by the mEXNEX$_c$ models. Across scenarios 3-6 both metrics simultaneously decrease for the independent model, and also demonstrates lowest percentage of correct inference compared to all other methods in scenarios 8-10. The mEXNEX$_{1/13}$ model has similar or lower percentage of correct inference in comparison to the EXNEX model but with consistently lower FWER values, while the mEXNEX$_0$ method has greater proportions of correct inference in scenario 3 compared to the standard EXNEX model (54.0% compared to 51.8%) but a 14% decrease under scenario 5. This reduction came with a 3.3% decrease in FWER. Under scenario 6, the methods shown to have higher power in Figure 2.3.2, also gave greater proportion of correct inference made across all baskets. Considering scenarios 8-10, the standard EXNEX model gives the best percentage of all correct inference with lower FWER than the BHM, CBHM and BMA approach in all cases. This is most prominent in scenario 9 where in 37.1% of simulation, the EXNEX model made correct conclusions in all 5

baskets, whereas, under the same scenario the $\text{mEXNEX}_{1/13}$ had a smaller value at 30.2% but with a 2% lower FWER.

In view of these results, when the sample size is fixed across baskets, the proposed mEXNEX_0 model controls error rates to a nominal level whilst also improving power over implementing an independent model. Improvements are also observed over the EXNEX model with consistently lower type I error rates but reduced power. Should interest lie more heavily on improving power over the control of error rates, the cut-off value for exclusion of heterogeneous baskets could be increased. Both cut-off values of 0 and 1/13 produce a model that either exceeds all other considered borrowing methods in performance or acts similarly to the standard EXNEX model.

Varying the True Response Rate Vector, p

There are an infinite number of data scenarios one could fall in when conducting clinical trial analysis, the scenarios listed in Table 2.3.1 are only a subset of these feasible cases. The data scenarios implemented above were selected to cover a wide range of cases, however, some important cases may not have been investigated.

To overcome this, a further simulation study was conducted within which, rather than fixing the true probability of success parameter prior to the study, for ever simulation run a new random truth vector, p, was generated with uniform probability across the ranges [0,0.15] and [0.35,0.5] (these ranges were set to ensure equal chances of lying in the null and non-null case respectively). Once p was generated, it was used to simulate data from a Binomial distribution. The goal of such a simulation study is to determine the operating characteristics on average over many different truth vectors in hope to capture what would occur in cases not investigated within the previous simulation study.

A total of 20,000 simulations for each borrowing method were run under the planned sample size case of 13 patients in each basket. Results are provided in Table 2.3.3,

with further descriptions and results for the realised sample size case provided in the Supporting Information (Section A.4).

Table 2.3.3: Operating characteristics of the information borrowing models in which the truth vector was randomly generated. This is conducted under the planned sample sizes.

Method	Type I Error Rate	Power	% All Correct	FWER
Independent	2.25	81.48	57.63	5.68
BHM	5.00	87.51	63.07	11.68
CBHM	2.27	79.21	54.21	5.48
BMA	4.51	86.87	62.57	10.62
EXNEX	3.15	86.01	63.64	7.86
mEXNEX_0	2.68	83.89	61.19	6.77
$\text{mEXNEX}_{1/13}$	3.16	85.97	63.78	7.80

Similar to the fixed scenario cases described above, the BHM and BMA have the highest error rates, but all methods have mean type I error rate less than the nominal 10% level. The reduced error rates come from, in some cases, the true response rate lying well below the null 15% level under which the Δ_α value was calibrated. The CBHM continues to behave similarly to an independent approach but with lower power.

The standard EXNEX model and $\text{mEXNEX}_{1/13}$ model behave very similarly in this study, both with type I error rate of 3.2% and power of 86.0%. This is not unexpected, as like in the previous study, when clusters of responses are present, the less conservative mEXNEX_c model begins to perform similarly to the standard EXNEX model due to it's inability to detect clusters of responses. When $c = 0$, error rates are far closer to the independent model at 2.7% (2.3% under an independent analysis) with 83.9% power, which although lower than the standard EXNEX model, is an increase of 2.4% over an independent analysis.

In terms of percentage of simulation runs in which the correct conclusion was made in all 5 of the baskets, both the standard EXNEX and $\text{mEXNEX}_{1/13}$ models have the highest value of around 63.7%. The BHM and BMA approach have similar but slightly smaller values compared to both methods but have 2.8-3.9% increase in FWER. The

mEXNEX$_0$ model gives both reduced percentage of all correct conclusions and FWER compared to all the aforementioned methods but does possess a 3.6% increase in all correct inference compared to an independent analysis.

To summarise, in the planned sample size case when the true response rate is varied, the BHM and BMA continue to display the most undesirable error rates whilst the independent analysis and CBHM lack power. The modified EXNEX model with $c = 1/13$ performs almost identically to the standard EXNEX model. When a more conservative cut-off value $c = 0$ is implemented, error rates are reduced by 0.5% compared to the standard EXNEX model but with a 2.1% reduction in power (but still a 2.4% improvement over an independent analysis.

2.3.2 Simulation Results: Realised Sample Sizes

Although the protocol planned for 13 patients per basket, 20, 10, 8, 18 and 7 patients were enrolled across the 5 baskets. The thresholds for efficacy, Δ_α, were calibrated based on the planned sample size of 13 per basket and was not re-calibrated based on these observed sample sizes. Similarly, the cut-off values c in the mEXNEX$_c$ model were not adjusted and were based on the planned equal sample size.

Percentage of rejection plots are provided in Figures 2.3.3 and 2.3.4 with full results in Tables A.1.4, A.1.5 and A.1.6 in the Supporting Information (Section A.1).

The calibration procedure for the CBHM needs more careful consideration here, as the previous calibration was based on equal sample sizes across baskets. A slight modification to step 2 of the process was made to cover all permutations of heterogeneity. Even with this adaption, when sample sizes are unequal, the calibrated values of a and b are much larger in magnitude than in the equal sample case. This leads to stronger borrowing where baskets are at least fairly homogeneous, producing much narrower posterior densities. These narrow posteriors, in some cases, have their mass lying entirely above q_0 and thus Δ_α is close to 1. This can cause a lack of power as it

Figure 2.3.3: Percentage of rejections of the null hypothesis for each information borrowing method under data scenarios 1-10 based on realised sample sizes of 20, 10, 8, 18 and 7 across the 5 baskets.

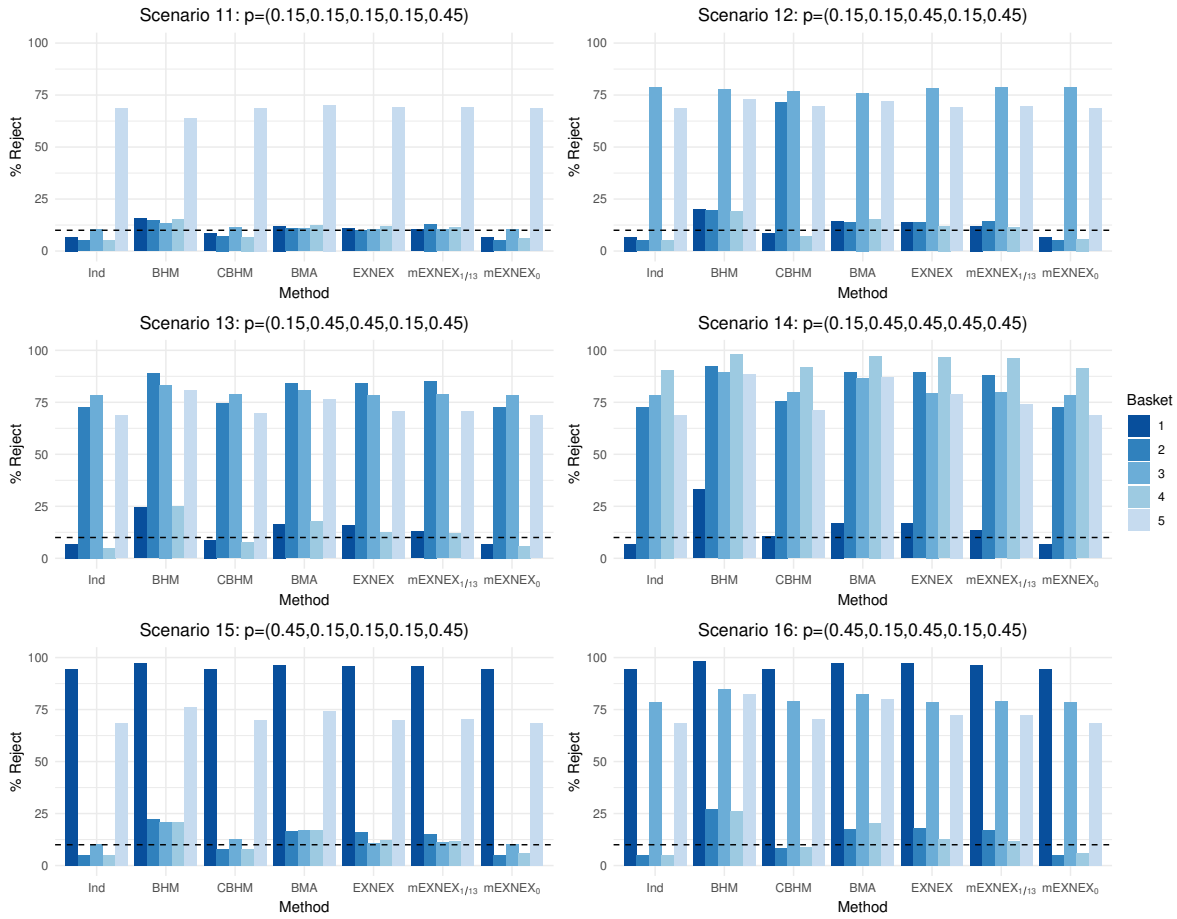


Figure 2.3.4: Percentage of rejections of the null hypothesis for each information borrowing method under data scenarios 11-16 based on realised sample sizes of 20, 10, 8, 18 and 7 across the 5 baskets.

makes it incredibly difficult to reject a hypothesis. To overcome this, we recommend calibrating a and b with the sample size fixed and equal for each basket at the averaged basket sample size. Analysis is then conducted using these tuned parameters with the observed unequal sample sizes. In this case, the average sample size across the baskets happens to be 13 patients per basket (with rounding) and thus, the a and b values used are the same as in the planned sample size case.

The type I error rate in scenario 1 lies below the nominal 10% level for all methods with the independent and mEXNEX_0 models giving the lowest values, whilst the BHM, BMA and EXNEX model are greater but are still approximately at or below 10%.

Under scenario 2, in which the first basket is effective to treatment, the BHM, EXNEX, $\text{mEXNEX}_{1/13}$ and BMA methods produce higher power than the independent analysis, at the cost of inflated error rates at 18.4%, 14.8%, 12.0% and 14.1% respectively. The CBHM and mEXNEX_0 model gives almost identical power values to the independent model (94.6% and 94.3% compared to 94.6%) but the mEXNEX_0 model gives error rates no greater than 10.5%. Similar results are also seen in scenario 7 in which the first basket is marginally effective. Now consider scenario 11 in which, like scenario 2 only one basket has an effective response rate but this is now the fifth basket as opposed to the first. This basket has a smaller sample size than that of the first basket at just 7 patients and thus, the power is uniformly lower for all methods, however, patterns of results remain the same in terms of method performance. Under this scenario, all methods (with the exception of the BHM) produce very similar power values ranging from 68.7%-70.2% but all have varying error rates. All borrowing methods have a higher error rate than that of an independent model, with inflation above the nominal 10% level present under the BHM, BMA, EXNEX and $\text{mEXNEX}_{1/13}$ model. The BHM has a much lower power in this case at 63.7%.

Across scenarios 3, 4 and 8, the $\text{mEXNEX}_{1/13}$ model gives similar/lower error rates compared to the EXNEX model and generally higher power values in baskets with

a small sample size i.e. baskets 2 and 3, with up to an increase of 2.4%. Similar power values are observed in basket 1 where the sample size is larger. The mEXNEX$_0$ model continues to control error rates at or below the 10% level but provides little to no improvement in terms of power over an independent approach. This is due to the conservative nature of this c value. When $c = 0$, under unequal sample sizes it is likely that all baskets will be treated as independent, as in the binary response setting, achieving identical response rates in baskets of different sizes is often impossible.

Scenario 5 again displays the improvement in power through using the borrowing techniques, with the exception of the mEXNEX$_0$ model for the aforementioned conservative nature. Ignoring the independent and mEXNEX$_0$ models for lack of power, the mEXNEX$_{1/13}$ model displays the lowest type I error rate of 17.9% which, although inflated, is considerably lower than the other borrowing methods, including the EXNEX model which has an error rate of 27.6%.

Similar to the planned sample size simulation, the BHM and BMA approach give greatest power in scenario 6 but at the cost of high error rates elsewhere. Across all baskets, the mEXNEX$_{1/13}$ model improves in power over the independent model by up to 16.51% but also at the cost of inflated error rates. However, this inflation occurs to a lesser extent than the EXNEX model across scenarios 2-5 with a maximum difference in error rates for the two methods at 9.7% which could be viewed as a highly significant margin.

Under scenarios 9 and 10, those baskets that have a marginally effective treatment effect show markedly improved power when information borrowing methods are implemented, with the mEXNEX$_{1/13}$ model obtaining up to a 20% improvement in power compared to an independent analysis under scenario 9 - note this comes with roughly a 3% inflation in error rate, but such inflation is less than the other borrowing methods.

Now consider the cases when the ordering of response rates is altered, under scenario 12 the two smallest baskets have the effective response rates whilst the larger baskets are

insensitive to treatment, the mEXNEX$_{1/13}$ model gives the greatest power for basket 3 at 78.5% (whilst the EXNEX model has power 78.1%) as well as improved power over the standard EXNEX model for basket 5 also (69.7% compared to 69.2%). This is alongside having a lower average type I error rate of 12.7% under mEXNEX$_{1/13}$ compared to 13.1% under the EXNEX model. In comparison to scenario 3, when the basket size is smaller, the performance of the BHM and a BMA approach worsens with higher errors and lower power values, whilst the performance of the mEXNEX$_{1/13}$ over other methods improves. The same conclusions can be drawn from scenarios 13-16 also.

Considering family-wise error rate and percentage of all correct conclusions across the 5 baskets, if the Δ_α values were calibrated to control FWER at 25% under the planned sample size and then applied to the realised sample size case, all methods give slightly inflated FWER values of over 25% under scenario 1 (see the Supporting Information, Section A, for full results). There is a 1-1 relation between low FWER and high percentage of cases where correct inference is made across all baskets with those showing the highest family-wise error rate also presenting lower percentages of correct inference.

The BHM and BMA approach give the highest FWER and lowest percentage of correct inference in all baskets across scenarios 2-16. Under scenarios 2 and 3 the mEXNEX$_{1/13}$ model has a FWER 5% smaller than the EXNEX model, producing similar values to the independent approach but with an improvement in power. Scenario 6 shows that models which typically inflate the error rate give the best proportions of correct inference across all baskets. The mEXNEX$_{1/13}$ model provides an increase of over 6% in comparison to an independent model. The percentage of all correct inference is smaller across scenarios where there are a few marginally effective baskets, i.e. scenarios 8-10 and this lines up with larger inflation in error rates.

Similarly to the planned sample size case, these results confirm that the choice of c value makes a big impact in the performance of the mEXNEX$_c$ model. The c values

were selected based on the planned sample size of 13 per basket and thus increments corresponding to 1 response were considered (i.e. 0, 1/13, 2/13,...), however, when sample sizes are unequal it would be beneficial to look at other potential values such as 0, 0.05, 0.1, etc.. A cut-off of 0.05 can be shown to perform well in this unequal sample size scenario, whereas the choice of $c = 0$ is far too conservative. In practice, when this occurs analysis can be conducted as specified in the trial protocol with the use of c based on planned sample sizes. Alternatively, one can re-calibrate based on the realised sample sizes and compare to original analysis to determine if there are any significant differences. It would be recommended to include instructions within the trial protocol on how to adjust the cut-off value for the mEXNEX$_c$ model once the realised sample sizes are known.

If the calibration of Δ_α accounted for unequal sample sizes, similar patterns in performance of each method is observed but with the impact of small sample sizes particularly evident. Results from a further simulation under the realised sample size with re-calibrated Δ_α based on the unequal nature are provided in the Supporting Information (Section A.2).

2.4 Analysis of VE-BASKET Results Using Information Borrowing Models

This section revisits the analysis of the VE-BASKET results using the described and proposed information borrowing methods. The data observed in the trial, and the posterior means for the response rate in each basket (and standard deviations) obtained by each method is given in Table 2.4.1. Also provided are the posterior probabilities of the response rate being greater than the null for each basket under each method, i.e. the decision making probability used at the conclusion of the trial. For this analysis, prior and parameter choices are provided in Table 2.3.2. Within the modified EXNEX

procedure, cut-off values, c, are chosen from $c = 0, 0.05, 0.1, 0.15, \ldots$. Through a simulation akin to that in Section 2.3, cut-off values of $c = 0.05$ and $c = 0.1$ were chosen, denoted $\text{mEXNEX}_{0.05}$ and $\text{mEXNEX}_{0.1}$ respectively.

For the EXNEX and mEXNEX_c models, specification of a prior probability vector, π, for assignment to the EX component is required. For each model, both the prior probability used and the posterior probabilities produced after model fit are listed below:

	Prior probability vectors:	Posterior probability vectors:
EXNEX:	$\pi = (0.50, 0.50, 0.50, 0.50, 0.50)$,	$\pi = (0.36, 0.50, 0.42, 0.39, 0.41)$.
$\text{mEXNEX}_{0.1}$:	$\pi = (0.74, 0.00, 0.00, 0.79, 0.74)$,	$\pi = (0.81, 0.00, 0.00, 0.85, 0.80)$.
$\text{mEXNEX}_{0.05}$:	$\pi = (0.00, 0.00, 0.00, 0.79, 0.79)$,	$\pi = (0.00, 0.00, 0.00, 0.74, 0.75)$.

The posterior probabilities for the EXNEX model decrease for all baskets compared to the prior values despite baskets 4 and 5 having homogeneous responses. In contrast, the $\text{mEXNEX}_{0.1}$ model increases between the prior and posterior probabilities which reflects the homogeneity of the response data. When $c = 0.05$, we observe a decrease in posterior probabilities from the prior values, however, they are still greater than in the EXNEX model, which suggests greater sensitivity to the presence of both homogeneous and heterogeneous baskets.

Table 2.4.1: Data summary of the VE-Basket trial with posterior means of the response rates obtained using the various information borrowing models alongside their standard deviations in brackets, as well as the posterior probability that the response rate is greater than the null.

Trial Data		NSCLC	Colorectal Cancer	Cholangiocarcinoma	ECD/LCH	Thyroid Cancer
Sample Size		20	10	8	18	7
ORR		0.40	0.00	0.13	0.33	0.29
Basket		1	2	3	4	5
Independent	$\hat{p}_k$	0.399 (0.11)	0.009 (0.03)	0.126 (0.11)	0.333 (0.11)	0.285 (0.16)
	$\mathbb{P}(p_k > 0.15\|D)$	0.996	0.008	0.325	0.968	0.777
BHM	$\hat{p}_k$	0.362 (0.10)	0.097 (0.09)	0.170 (0.11)	0.309 (0.10)	0.267 (0.13)
	$\mathbb{P}(p_k > 0.15\|D)$	0.994	0.259	0.518	0.966	0.809
CBHM	$\hat{p}_k$	0.398 (0.11)	0.012 (0.03)	0.125 (0.11)	0.331 (0.11)	0.281 (0.16)
	$\mathbb{P}(p_k > 0.15\|D)$	0.996	0.012	0.320	0.970	0.770
BMA	$\hat{p}_k$	0.368 (0.09)	0.058 (0.08)	0.213 (0.09)	0.331 (0.09)	0.309 (0.12)
	$\mathbb{P}(p_k > 0.15\|D)$	0.997	0.120	0.648	0.981	0.899
EXNEX	$\hat{p}_k$	0.384 (0.10)	0.059 (0.07)	0.171 (0.12)	0.326 (0.10)	0.288 (0.14)
	$\mathbb{P}(p_k > 0.15\|D)$	0.996	0.113	0.501	0.971	0.825
$\text{mEXNEX}_{0.1}$	$\hat{p}_k$	0.384 (0.10)	0.061 (0.06)	0.162 (0.11)	0.338 (0.10)	0.318 (0.13)
	$\mathbb{P}(p_k > 0.15\|D)$	0.997	0.089	0.454	0.983	0.904
$\text{mEXNEX}_{0.05}$	$\hat{p}_k$	0.398 (0.10)	0.061 (0.06)	0.162 (0.11)	0.328 (0.10)	0.301 (0.14)
	$\mathbb{P}(p_k > 0.15\|D)$	0.996	0.088	0.455	0.973	0.857

The mEXNEX$_{0.05}$ model, only allows borrowing between baskets 4 and 5 with probability 0.79. This results in standard deviations lower in these baskets compared to the independent model. When $c = 0.1$, the NSCLC basket is now included in the borrowing component with probability 0.74. This results in the estimated response rate in the first basket being pulled down as information is borrowed from baskets 4 and 5. The estimates and standard deviations for baskets 2 and 3 are identical for both c values as they are assigned to the NEX component. The mEXNEX$_{0.1}$ model has marginally smaller standard deviations compared to the EXNEX model with similar point estimates.

The results in Table 2.4.1 also demonstrate that using the independent model on baskets with small sample sizes leads to estimates with less precision due to the lack of borrowing. The CBHM results match that of the independent model due to the 'strong' definition of heterogeneity in it's calibration procedure. There is clear heterogeneity between basket's 1 and 2 in which the ORR is 0.4 and 0 respectively and thus the CBHM treats all baskets as being independent with $\sigma^2 \approx 383$.

The estimates using the BHM are pulled towards the common mean so the values are different to the ORR values, this is most evident in the second basket where the BHM estimates $\hat{p}_2 = 0.1$ while the ORR is 0. This is a direct result of the pull towards the common mean. A similar pattern is observed under the BMA method as the averaging procedure puts some weight on models that borrow between all baskets despite heterogeneity.

Focusing on the posterior probabilities of exceeding the null response rates, all methods give similar values for basket 1 which has a larger sample size and ORR value. This will likely lead to the treatment being deemed effective in basket 1 regardless of the method. However, the same can't be said for basket 2 in which these probabilities vary across all methods, giving a value of approximately 0.01 under a stratified analysis, compared to 0.25 under the BHM. This could lead to potentially differing conclusions

regarding the efficacy of a treatment based on the method used to analyse the results. Methods that borrow information between all baskets tend to have higher posterior probabilities when basket sample size is small compared to an independent analysis and methods such as the CBHM and mEXNEX$_c$ which borrow information to a lesser extent.

These results highlight that, as expected, the choice of borrowing method can impact inference made at the conclusion of a trial, especially in the case of heterogeneity across baskets. Heterogeneity causes a pull towards the common mean under most borrowing methods resulting in estimates different to the ORR values whilst having an even bigger impact on the decision probabilities used at the conclusion of the trial. However, the results also demonstrate benefits of borrowing in terms of increase in precision of point estimates, particularly when the sample size is small such as in the Thyroid cancer basket which has just 7 patients. From these differences in results, we would promote careful planning and pre-trial evaluations to ensure that the borrowing method used is appropriate for the study.

2.5 Discussion

Presented here were several Bayesian information borrowing techniques within a basket trial setting, alongside a proposed modification to the EXNEX model. Through simulation, the BHM, EXNEX model and a BMA approach were shown to have inflated error rates in the presence of baskets with heterogeneous response rates, while the CBHM lacks power in such a scenario.

Exploration of the methods applied to unequal sample sizes across baskets highlighted the inadequacy of the current calibration procedure in the CBHM which only previously considered equal sample sizes across baskets. A generalisation of this calibration is made to handle the presence of unequal sample sizes, a situation that commonly

arises in the clinical setting.

The proposed method has been shown to improve error control while increasing power over an independent analysis. This proposed method is robust to the presence of a heterogeneous basket as it is able to identify its difference in response and thus does not borrow information from it, while still retaining borrowing between homogeneous baskets with a probability determined by similarity in response through Hellinger distances.

The use of Hellinger distances has already been proposed for use in information borrowing in the basket trial setting by Zheng and Wason (2022). However, they utilise the metric on data with continuous endpoints and a control arm, to stipulate a commensurate prior based on pairwise Hellinger distances. The mEXNEX$_c$ model uses averaged Hellinger distances to compute the prior probability of borrowing within the EXNEX model. Alternative distance metrics were considered but were shown to have less error control to that proposed in this chapter and are hence omitted.

The mEXNEX$_c$ model has been specified as a two-step procedure, within which we first remove heterogeneous baskets to treat as independent and then utilise these Hellinger distances to specify the prior borrowing probabilities between the remaining baskets. In Section A.5 of the Supporting Information, explanation is provided as to why both of these steps are utilised in place of making just one of these modifications. Justifications are provided based on several thorough simulation studies, the first of which explored the performance of the one 1-step vs. 2-step methods under the simulation setting outlined in Section 2.3 which highlighted the need for the first step - i.e. removal of heterogeneous baskets - in order to control the type I error rate. We then continued exploration of the differences in approaches through a further simulation study that varied one design parameter at a time, i.e. changed the number of baskets (of which further simulation studies under $K = 3$ and $K = 10$ baskets are presented in Section A.6), changed the sample size or changed the target response rate. From this

we concluded that the 2-step mEXNEX$_c$ model as proposed in this chapter performs more favourably over a 1-step modified EXNEX model when the sample size is very small or large, when we have a smaller number of baskets and when the target response rate is closer to the null response rate. This is a more realistic trial setting and hence why the 2-step mEXNEX$_c$ model has been proposed, although an argument could be made in some cases to use just a 1-step procedure in which heterogeneous baskets are removed and the remaining borrowing probabilities are fixed at 0.5.

The performance of the modified EXNEX model is reliant on the cut-off specification for assigning a basket for independent analysis, which is selected to balance the trade-off between power improvement and control of type I error rate. When chosen to favour power improvement, the proposed method reduces error rates in the presence of a single heterogeneous basket and improves power when all baskets are sensitive to treatment. However, when clusters of responses are observed, the proposed method increases the probability of borrowing between all baskets and hence error rates increase and the method performs similarly to the standard EXNEX model. Whereas, if the cut-off is chosen to control error rates this inflation is not present across any of the simulation scenarios considered and power is improved in comparison to an independent analysis. As a result, implementing this newly proposed modified EXNEX model with a suitable cut-off value, produces a model that either exceeds all other borrowing methods considered here in terms of performance or acts similarly to the standard EXNEX model.

A draw towards the standard EXNEX model is it's ability to borrow between multiple subsets of baskets by incorporating more than one exchangeability component in its mixture distribution in Model 2.2.4. The mEXNEX$_c$ model could benefit from extension to allow for this feature. This would lead to better handling of borrowing within clusters of homogeneous responses.

Other alternative approaches for information borrowing in the basket trial setting

are outlined in the literature, these include the MUCE design (Lyu et al., 2023), Liu's two-path approach (Liu et al., 2017) and the RoBoT design (Zhou and Ji, 2020) to name a few. Comparisons between the proposed mEXNEX$_c$ model and the above methods have not yet been made.

Adaptive design features such as interim analyses with futility/efficacy stopping are desirable in most clinical trials and has been considered in the work by Jin et al. (2020), Berry et al. (2013), Chu and Yuan (2018) and Psioda et al. (2021). However, no such adaptive design features were considered in this chapter which could be considered a limitation. The methodology described here could be extended to incorporate such features and future work into this aspect is being conducted.

2.6 Appendix

2.6.1 Simulation Prior and Parameter Specification

For the simulation study in Section 2.3, priors are chosen to match those suggested in the models literature. The following priors are used for the Simulation study:

- Independent model:

$$\begin{aligned} Y_k &\sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K \\ \theta_k &= \log\left(\frac{p_k}{1-p_k}\right), \\ \theta_k &\sim \text{N}(\text{logit}(0.15), 10^2), \end{aligned}$$

- Bayesian Hierarchical model:

$$\begin{aligned} Y_k &\sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K \\ \theta_k &= \log\left(\frac{p_k}{1-p_k}\right) \sim N(\mu, \sigma^2), \\ \mu &\sim N(\text{logit}(0.15), 10^2), \\ \sigma &\sim \text{Half-Cauchy}(0, 25). \end{aligned}$$

- Calibrated Bayesian hierarchical model:

$$\begin{aligned} Y_k &\sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K \\ \theta_k &= \log\left(\frac{p_k}{1-p_k}\right) \sim N(\mu, \sigma^2), \\ \mu &\sim N(\text{logit}(0.15), 10^2), \\ \sigma^2 &= \exp\{a + b\log(T)\}, \end{aligned}$$

  where, through tuning, $a = -7.25$ and $b = 5.86$ based on a sample size of 13 per basket. The chi-squared test statistic is used to compute T as in Equation 2.2.3.

- Bayesian model averaging: A weakly informative Beta prior is placed on the response rates so $p_{S_j}|\mathcal{M}_j \sim \text{Beta}(a_0, b_0)$ where $a_0 = q_1 = 0.45$ and $b_0 = 1 - q_1 = 0.55$. The prior $f(\mathcal{M}_|) \sim P_j^2$ is placed on the models, where P_j is the number of distinct response rates in model j.

- EXNEX: For the standard EXNEX model equal prior mixture weights for the EX/NEX components are used and thus $\pi_k = 0.5$ for all k baskets. A plausible guess of the true response rate is chosen to be $\rho_k = 0.35$ (a value that is considered low but still indicative of a response) for all k baskets:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && M_{1k} \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)} \\
&\theta_k = \log\left(\frac{p_k}{1-p_k}\right), && \mu \sim \text{N}(\text{logit}(0.15), 10^2), \\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && \sigma \sim \text{Half-Normal}(0, 1), \\
&\delta_k \sim \text{Bernoulli}(0.5), && M_{2k} \sim N(-0.62, 4.4^2), \quad \text{(NEX)}
\end{aligned}
$$

with the parameters of the NEX component computed through the following Neuenschwander et al. (2016):

$$
m_k = \log\left(\frac{\rho_k}{1-\rho_k}\right), \quad \nu_k = \frac{1}{\rho_k} + \frac{1}{1-\rho_k}. \tag{2.6.1}
$$

- Modified EXNEX: The same structure and prior choices as the standard EXNEX model with the exception of the prior on σ. Rather than applying the prior $\sigma \sim \text{Half-Normal}(0, 1)$ the prior is placed on σ^2, i.e. $\sigma^2 \sim \text{Half-Normal}(0, 1)$. The mixture weights have a Bernoulli prior with prior parameter of success, π_k, which are calculated via the Hellinger distance with cut-off c chosen to be 0 and 1/13.

Chapter 3

How to Add Baskets to an Ongoing Basket Trial with Information Borrowing

3.1 Introduction

In the oncology setting, significant research into cancer genomics and understanding the underlying genetic cause of disease has catapulted the field of personalised medicine, within which treatments are targeted to a specific genetic makeup, to the forefront of clinical trial design (Simon and Roychowdhury, 2013). This shift away from disease specific treatments towards genetically targeted treatments has led to the development of basket clinical trials.

Basket trials are a form of master protocol in which a single treatment is administered to patients across different disease types, all of whom possess the same genetic aberration. Different disease type sub-populations form their own treatment basket (Sargent and Renfro, 2017). Typically, basket trials are implemented in early stages of the drug development process in order to determine if a treatment is efficacious against

each of the individual baskets on the trial (Park et al., 2019). They often consist of a single treatment arm using a small number of patients.

One of the main benefits of basket trials is that they allow for testing of treatments on rare diseases that would not traditionally warrant their own investigation due to their limited sample size (Chu and Yuan, 2018) and financial and time constraints. By allowing for testing on multiple disease types in a single study, the drug development process is substantially expedited. These basket trials also provide flexibility by utilising adaptive design features, which allow for modification of the design and analysis while the study is still ongoing. Such modifications include interim analysis with futility and efficacy stopping, sample size adjustment, or as is the focus of this work, the addition of a single or multiple baskets to an ongoing trial. This situation arises when it is identified that a new group of patients may benefit from the treatment, where these patients harbour the genetic aberration under investigation but suffer from a different disease type.

Several prominent clinical trials have utilised the addition of a treatment arm/basket. The VE-BASKET trial (Hyman et al., 2015), exploring the effect of Vemurafeib on cancers with the BRAFV600 mutation, is an example of such a study in which two new baskets were formed of patients on the trial due to sufficient accrual rates of patients of two cancer types. Likewise, the basket trial looking at the treatment of tucatinib and trastuzab on solid tumours with the HER2 alteration (Reck et al., 2021) allowed the opening of disease-specific cohorts during the trial as a part of the trial protocol. These examples act as motivation of the work presented in this chapter, with the purpose to explore methodology for handling such additions of treatment groups.

Although basket trials are desirable as they allow the testing of rare diseases, this does introduce challenges in cases where sample sizes are limited. In such situations, issues such as lack of statistical power and precision of estimates arise. This can be amplified in baskets that are added part-way through an ongoing trial. The combination

of reduced recruitment rate (when the new disease type is rare) and shorter recruitment time due to the late addition to the trial, can result in a further reduction in sample sizes compared to baskets that opened at the beginning of the trial. To tackle the issue of small sample sizes within baskets, Bayesian information borrowing methods were proposed for use in basket trials. These methods utilise the assumption that, as patients across baskets share the same genetic mutation, they will have a similar response to the treatment. As such, patients are 'exchangeable' between baskets, meaning patients can be moved between treatment baskets without changing the overall treatment effect estimates (Oakes, 2013). One can use this assumption to draw on information from one basket when making inference in another. This has the potential to improve power and precision of estimates, particularly in the presence of small sample sizes. However, when the exchangeability assumption is violated, and there is heterogeneity amongst baskets' responses, any information borrowing has the potential to inflate the type I error rate. This trade-off between power improvement and error rate inflation amongst heterogeneous baskets is a well known issue and has been observed in several simulation studies throughout the literature including that by Chu and Yuan (2018), Jin et al. (2020) and the work outlined in Chapter 2.

Over recent years, several prominent methods for information borrowing in basket trials have been proposed. These include the Bayesian hierarchical model (BHM, Berry et al., 2013) and several adaptations to this method, such as the calibrated Bayesian hierarchical model (CBHM, Chu and Yuan, 2018) which defines the prior on the borrowing parameter as a function of homogeneity, the exchangeability-nonexchangeability model (EXNEX, Neuenschwander et al., 2016) which allows for flexible borrowing between subsets of baskets and the modified exchangeability-nonexchangeability model outlined in Chapter 2 which modifies the EXNEX model to account for homogeneity/heterogeneity between baskets.

This chapter proposes and investigates different approaches for the analysis of newly

added baskets under an information borrowing structure, which primarily utilises the EXNEX model. To identify when and which approach is deemed appropriate for use, thorough simulation studies under a variety of setting were conducted, monitoring the type I error rate and power. The simplest approach to such an addition would be to analyse the new baskets akin to baskets that were already in the trial at the start, a problem which is mathematically equivalent to a case of unequal sample sizes. This chapter also explores additional methodology, motivated by the concern that new baskets could negatively impact the type I error rate and power of existing baskets should results be heterogeneous. However, substantial power can be gained by borrowing from new baskets in cases of homogeneity. Error control in the the new basket must also be considered. Results of thorough simulation studies are provided to compare such approaches in order to identify when and how it can be beneficial to add a new basket to an ongoing trial as opposed to running a separate investigation for the new basket(s).

The second novel aspect of this chapter regards the calibration of efficacy criteria. When implementing Bayesian borrowing models, posterior probabilities are computed and compared to some pre-defined cut-off value in order to determine whether or not a treatment is efficacious in each of the baskets. Traditionally, these cut-off values are calibrated through simulation studies under a global null scenario, where all baskets have a truly ineffective response rate. This calibration aims to control the basket specific type I error rate to a nominal level. However, when the cut-off value is applied to cases where at least one basket is non-null, it is not guaranteed that error rates will remain controlled at the nominal level when information borrowing is utilised (Kopp-Schneider et al., 2020). In fact, inflation in error rates often occurs in cases of heterogeneity, as borrowing information causes shifts in the posterior probabilities away from the true treatment effect. This brings into question whether calibrating under the global null is sufficient, as more often than not, there is an expectation that the treatment is efficacious in at least one basket. In this chapter we propose a novel calibration

technique, called the **R**obust **Ca**libration **P**rocedure (RCaP), which controls the type I error rate *on average* across several possible true response rate data scenarios, with the potential to weight scenarios based on importance and prior likelihood of occurring in the trial. Presented in this chapter is a comparison between operating characteristics under the traditional approach of calibrating under the global null and under this novel procedure.

This chapter is structured as follows, we begin in Section 3.1.1 with introducing a motivating example, the VE-BASKET study. In Section 3.2 we then describe analysis models, approaches for the analysis of newly added baskets, and outline the novel calibration procedure, RCaP. Results of several simulation are presented in Section 3.3 starting with a comparison of calibration techniques, followed by results of simulation studies to compare performance of approaches for the addition of a newly identified basket.

3.1.1 Motivating Trial: The VE-BASKET Study

The VE-BASKET trial was a phase II basket trial, investigating the effect of Vemurafeib on several cancer types possessing the BRAFV600 mutation (Hyman et al., 2015). A total of 122 patients were enrolled across all baskets, with efficacy evaluated after eight weeks of treatment. The primary endpoint was the overall response rate (ORR) with a null response rate of 15% indicating inactivity and target response rate of 45%. A response rate of 35% was considered low but still indicative of a response. Sample sizes of 13 patients per basket were obtained through a Simon's two stage design (Simon, 1989) based on 80% power and 10% type I error rate.

The trial opened with six disease specific baskets: non-small-cell lung cancer (NSCLC), ovarian cancer, colorectal cancer, cholangiocarcinoma, breast cancer and multiple myeloma. Also present was an 'all other' basket consisting of patients with different disease types with the BRAFV600 mutation. This initial trial structure was adapted based on re-

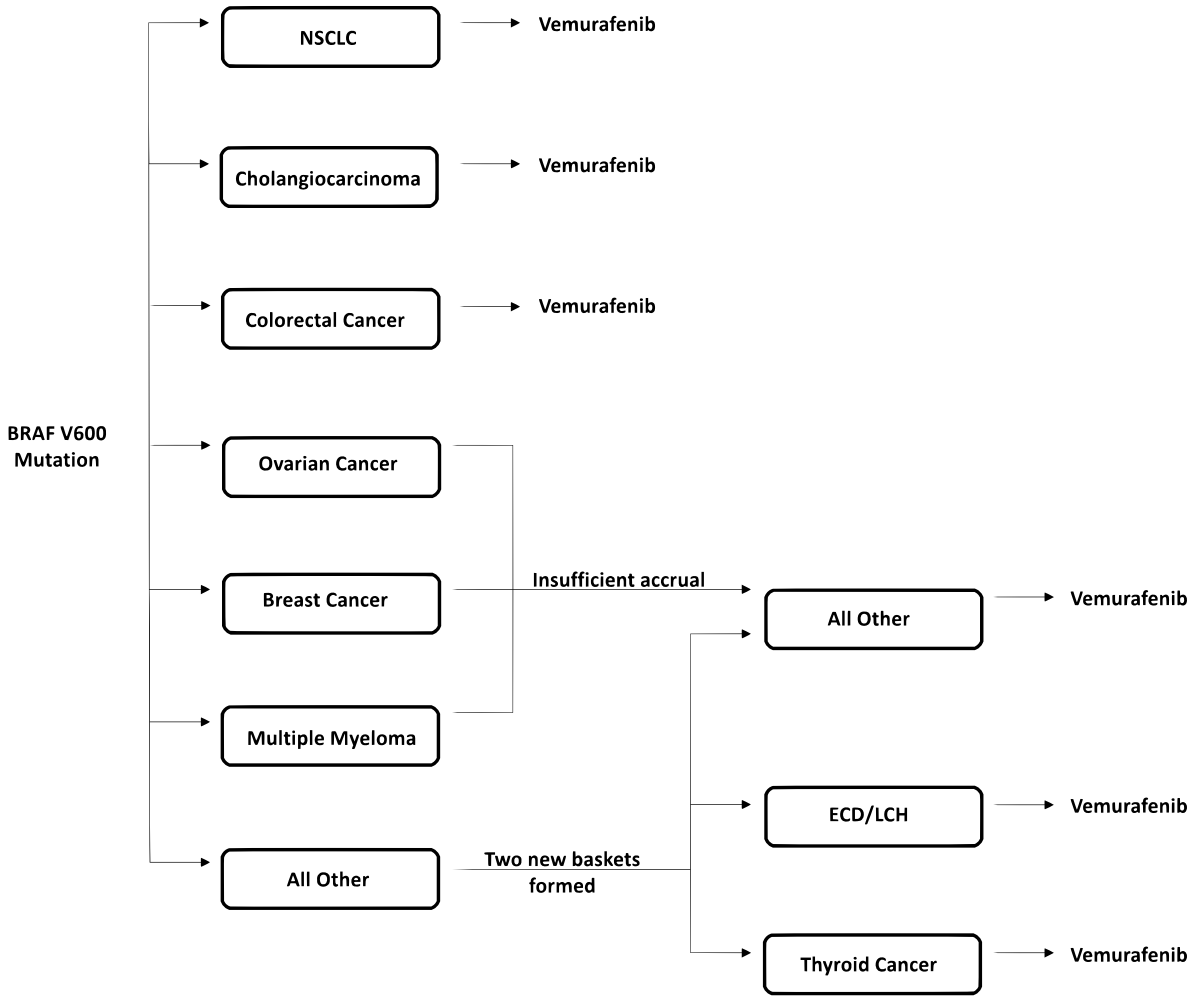


Figure 3.1.1: VE-BASKET Trial Design. Vemurafenib is tested on several cancer types, with two new baskets formed from the 'all other' group in the trial.

cruitment rates, with the breast cancer, ovarian cancer and multiple myeloma baskets closing due to insufficient accrual. Patients were moved from these baskets to the 'all other' basket for analysis. Due to sufficient number of patients in the 'all other' group, two new baskets were formed and added to the trial: an Edrheim-Chester disease or Langerhans'-cell histiocytosis (ECD/LCH) basket and a anaplastic thyroid cancer basket. Figure 3.1.1 displays the general trial schematic.

This highlights the flexible nature of a basket trial. Although not newly identified baskets, with new disease groups forming from patients already on the trial, the VE-BASKET trial demonstrates the addition of baskets, hence bringing about the question

on how to conduct analysis in such a setting.

3.2 Methodology

3.2.1 Setting

This chapter focuses on a single treatment arm within each basket and considers binary endpoints, in which a patient either responds positively to a treatment or does not. Consider a basket trial with a total of K baskets. Denote the number of responses in basket k by Y_k, which follows a binomial distribution, $Y_k \sim \text{Binomial}(n_k, p_k)$, with n_k and p_k indicating the sample size and response rate in basket k. Interest lies in estimating the unknown response rate p_k. Denote q_0 and q_1 as the null and target response rate respectively.

Now consider a case where baskets of patients are added to an ongoing trial and thus split the K baskets into two sets. Let K_0 be the total number of existing baskets that began the trial, thus having $K' = K - K_0$ new baskets added part way through the study. Existing baskets are indexed $k_0 = 1, \ldots, K_0$ and new baskets $k' = K_0 + 1, \ldots, K$. Note that a new basket, k', may be added at any time during the study and it is not required that all new baskets be added at the same time.

The objective is to test the family of hypotheses:

$$\begin{aligned} &H_0 \,:\, p_{k_0} \leq q_0 \qquad vs. \qquad H_a \,:\, p_{k_0} > q_0, \qquad k_0 = 1, \ldots, K_0, \\ &H_0 \,:\, p_{k'} \leq q_0 \qquad vs. \qquad H_a \,:\, p_{k'} > q_0, \qquad k' = K_0 + 1, \ldots, K. \end{aligned}$$

To test these hypotheses, a Bayesian framework is utilised. Posterior probabilities are used to determine the efficacy of the treatment on each of the individual baskets in the trial. As such, given observed response data D, the treatment is deemed effective in an existing basket k_0 if $\mathbb{P}(p_{k_0} > q_0 | D) > \Delta_{k_0}$ and effective in a new basket k' if

$\mathbb{P}(p_{k'} > q_0|D) > \Delta_{k'}$. Both cut-off values Δ_{k_0} and $\Delta_{k'}$ are typically determined through *calibration* in order to control some metric, often related to false decision making, at a nominal level. Traditionally this calibration is done under a global null scenario in which all baskets are ineffective to treatment, in order to control the basket-specific type I error rate to a nominal level (Kaizer et al., 2022; Hobbs and Landin, 2018; Jin et al., 2020).

Note that calibration of these cut-off values mostly occurs prior to the trial commencing, and hence before observed sample sizes are known. Due to this uncertainty, assumptions must be made for the sample sizes in both existing and new baskets in order to conduct calibration. Should the impact of much greater or much smaller sample sizes than planned be of concern, one could calibrate based on the 'worst case scenario' for the sample sizes (i.e. the sample size which is expected to observe the greatest type I error rate for instance).

3.2.2 The Exchangeability-Nonexchangeability Model

Information borrowing models utilise the exchangeability assumption, which states that as patients across all baskets share a common genetic component, their response to treatment will be similar. Thus information can be shared between baskets in order to improve inference. The Bayesian hierarchical model (BHM) first outlined by Berry et al. (2013) is a key basis for many information borrowing models, one of which is the exchangeability-nonexchangeability (EXNEX) model proposed by Neuenschwander et al. (2016). The EXNEX model consists of two components:

1. EX (exchangeable component): with prior probability π_k, basket k is exchangeable and a Bayesian hierarchical model is applied. Information borrowing is therefore conducted between all baskets assigned to the exchangeable component.

2. NEX (nonexchangeable component): with prior probability $1 - \pi_k$, basket k is

nonexchangeable with any other basket, and as a result is analysed independently.

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1, \ldots, K && M_{1k} \sim \text{N}(\mu, \sigma^2), && \text{(EX)}\\
&\theta_k = \log\left(\frac{p_k}{1-p_k}\right), && \mu \sim \text{N}(\text{logit}(q_0), \nu_\mu),\\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && \sigma \sim g(\cdot),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && M_{2k} \sim \text{N}(m_k, \nu_k). && \text{(NEX)}
\end{aligned} \tag{3.2.1}
$$

As outlined in the model specification (3.2.1), the EX component has the form of a BHM with the log-odds of the response rates for each basket following a normal distribution, centred around a common mean μ with variance σ^2. Borrowing occurs between baskets in the EX component where estimates of response rates are shrunk towards the common mean μ with the degree of shrinkage controlled by σ^2. As σ^2 tends to zero, borrowing moves towards complete pooling of results, however, as it tends to infinity a stratified analysis is conducted on each basket. The prior on μ is centred around the average null response rate across the baskets with a large variance, whilst the prior on σ, $g(\cdot)$, is more widely debated with Inverse-Gamma, Half-Normal or Half-Cauchy priors implemented across the literature (Gelman, 2006). It is suggested that a Half-Normal(0,1) prior is to be placed on σ as this allows for anywhere between a small and very large amount of heterogeneity between baskets (Neuenschwander et al., 2016).

Issues arise in a BHM when the exchangeability assumption is violated, which occurs in the presence of heterogeneous baskets. In such cases, when information is borrowed between all baskets, the type I error rate is likely to inflate as the posterior probabilities are pulled towards the common mean, μ, and away from the true treatment effect. The EXNEX model relaxes the full exchangeability assumption, allowing for some heterogeneity between treatment effects (thereby reducing type I error rate inflation) through the incorporation of the NEX component within which baskets are analysed independently, with basket-specific priors on the logit transformed response

rates. Neuenschwander et al. (2016) propose setting the parameters as follows:

$$m_k = \log\left(\frac{\rho_k}{1-\rho_k}\right), \quad \nu_k = \frac{1}{\rho_k} + \frac{1}{1-\rho_k}, \tag{3.2.2}$$

where ρ_k is a plausible guess for the true response rate in basket k.

The prior probabilities, π_k, for assignment to the EX/NEX component are selected prior to the trial. There is often little to no information available on the probability of exchangeability of baskets before the trial, so it is suggested to fix $\pi_k = 0.5$ for all k baskets. Alternatively, a Dirichlet prior could be placed on these values, however, Neuenschwander et al. (2023) prove that only the mean of the weight distribution affects inference in this case.

3.2.3 Approaches for Adding A Basket

We now propose four different approaches for the calibration and analysis of newly added baskets to an ongoing basket trial. In all four cases existing baskets are analysed through an EXNEX model, however, treatment of the new basket varies. Approaches are outlined below, as well as, summarised in Table 3.2.1.

1. **IND** - INDependent analysis of the new basket.

   Analyse the K_0 existing baskets by applying an EXNEX model (as in Model (3.2.1)) and calibrate Δ_{k_0} based on the same model. Analyse the K' new baskets independently (modelled as in the NEX component in Model (3.2.1)) to existing baskets and calibrate $\Delta_{k'}$ based on the same model.

   Analysing the new basket as independent may be considered desirable as it eliminates potential negative effects of smaller sample sizes in new baskets on inference in existing baskets.

2. **UNPL** - UNPLanned addition of a new basket.

Calibrate Δ_{k_0} based on an EXNEX model applied to the K_0 existing baskets. When conducting analysis borrow between all K baskets through an EXNEX model. For cases of equal sample size set $\Delta_{k'} = \Delta_{k_0}$ for the new basket. If the sample size is unequal for the K' new baskets compared to the K existing baskets, set $\Delta_{k'} = \Delta_{i_0}$ where existing basket i has sample size n_i closest to the sample size of the new basket k', i.e. $i = \arg\min_i\{|n_i - n_{k'}|\}$.

This is a naive analysis in that cut-off values are not adjusted despite the additional baskets. This may occur when an addition is not planned for, but once it occurs it is believed that borrowing from a new basket will improve inference for both existing *and* new baskets due to the extra information gained.

3. **PL1** - PLanned addition of a new basket in which a single EXNEX model is applied.

   Calibrate Δ_{k_0} and $\Delta_{k'}$ assuming that new baskets will be added during the study. To calibrate and analyse, borrow between all K baskets (new and existing) through an EXNEX model. Effectively, this equates to calibration under unequal sample sizes and has two subsets:

   (a) The time of addition of the new basket(s) is known. In this case, the sample sizes, n_k, for each of the $k = 1, \ldots, K$ baskets are known and fixed in the calibration procedure. The new baskets are assumed to have equal recruitment rates to the existing baskets unless evidence exists to the contrary.

   (b) The time of addition of the new basket(s) is unknown. In this case further simulation studies are required to explore the effect of sample size on operating characteristics. Based on these exploratory simulation studies, it may be desirable to calibrate based on the sample size which produced the highest type I error rate inflation. Utilising this cut-off value will ensure better error control, however may come at the cost of reduced power if overly

conservative.

The situation in which it is known for certain that new baskets will be added may occur if it is apparent that a basket of patients will benefit from the study, however, is not ready in time for the commencement of the trial, whether this is due to logistical issues or diagnostic techniques or some other factor. Thus it is planned to add the basket at a later time. This time of addition may be fixed as in PL1(a) but it may be desirable to add a basket as soon as it is available, thus falling into the case of PL1(b) when timing of addition is unknown.

4. **PL2** - PLanned addition of a new basket in which two EXNEX models are applied.

   Calibrate Δ_{k_0} based on an EXNEX model applied to just the K_0 existing baskets and thus, when analysing the existing baskets, do not borrow from any new baskets. Calibrate $\Delta_{k'}$ based on an EXNEX model applied to all K baskets. Therefore, when analysing new baskets, information is borrowed between all baskets, new and existing. This results in two EXNEX models and, like PL1, consists of two subsets: (a) Timing of addition is known and fixed and (b) Timing of addition is unknown.

   As in IND, analysing baskets in this way will eliminate the effect of reduced sample sizes in new baskets on estimation of response rates in existing baskets. However, by allowing full information borrowing between all baskets when analysing the new baskets, one may combat the issue of lack of statistical power and precision of estimates that arises due to the limited sample size.

Under both IND and PL2, the calibration and analysis for existing baskets are equivalent, with an EXNEX model applied to all K_0 existing baskets, independent of any new baskets. Similarly, for new baskets, PL1 and PL2 are equivalent as under both, when calibrating and analysing any of the K' new baskets, information is borrowed between all K baskets in the trial by fitting an EXNEX model. Full model specifications

Table 3.2.1: Summary of approaches for analysis and calibration when adding a basket where k_0 denotes existing baskets and k' denotes new baskets.

Approach	**Calibration**		**Analysis**	
	Δ_{k_0}	$\Delta_{k'}$	Existing Baskets	New Baskets
IND	EXNEX on all k_0	Independent on all k'	EXNEX on all k_0	Independent on all k'
UNPL	EXNEX on all k_0	$\Delta_{k_0} = \Delta_{k'}$	EXNEX on all k	
PL1	EXNEX on all k		EXNEX on all k	
PL2	EXNEX on all k_0	EXNEX on all k	EXNEX on all k_0	EXNEX on all k

and a further table summary are provided in Table 3.5.1 in Appendix 3.5.

3.2.4 RCaP: Robust Calibration Procedure

A treatment is deemed effective in basket k if the posterior probability that the response rate, p_k, is greater than q_0, exceeds a cut-off value Δ_k. In a few basket trial cases, such as the work by Zheng and Wason (2022) and Ouma et al. (2022a), these Δ_k values are fixed at some value, i.e. 0.975, however, an alternative is to calibrate the cut-off value in order to control some operating characteristic to a desirable level.

This was implemented by Kaizer et al. (2022), Hobbs and Landin (2018), Chu and Yuan (2018), Jin et al. (2020) and Berry et al. (2013), who followed a conventional approach where Δ_k was calibrated under a single simulation scenario (i.e. a vector of probabilities that reflect response rates in each of the baskets in the trial), typically this is the global null scenario in which the treatment is ineffective across all baskets. In each of these cases Δ_k was calibrated to achieve an $100\alpha\%$ type I error rate in each basket under the global null. However, this type of calibration does not guarantee error rate control across other scenarios when information borrowing is implemented. When borrowing information from heterogeneous and effective baskets, the posterior probabilities are pulled upwards for baskets with an ineffective response rate compared to that under the global null scenario, thus increasing the probability of exceeding the calibrated value, Δ_k. Therefore, error control is only guaranteed in the global scenario in which Δ_k was calibrated under, with other scenarios likely to demonstrate undesirable

error rate inflation. This is observed in the simulation study conducted in Chapter 2, with the EXNEX model, in the worst case, producing a relative increase in type I error rate of 72.5% compared to the nominal 10% level. Similar findings are presented by Jin et al. (2020) (236% relative increase) and Chen and Hsiao (2023) (135% relative increase). Although Δ_k is typically calibrated to control the type I error rate, the calibration procedure remains the same for the control of any metric obtained from the posterior density such as the family-wise error rate or power.

We propose a novel calibration procedure, the Robust Calibration Procedure (RCaP), where as opposed to calibrating under a single global null scenario (which we refer to as the 'calibration under the global null approach'), Δ_k is calibrated across numerous potential scenarios so that some metric, Q, is controlled *on average* across potential trial outcomes. Simulation scenarios may be weighted in importance by their probability of occurring in the trial, or if no information is provided on potential outcomes, scenarios can be equally weighted.

Consider a case with M simulation scenarios $\boldsymbol{p_1}, \ldots, \boldsymbol{p}_M$ one wishes to calibrate across. Denote the sample size and true response rate of basket k under scenario m as n_{mk} and p_{mk} respectively with $k = 1, \ldots, K$ and $m = 1, \ldots, M$. The simulation scenarios are represented by vectors consisting of these true response rate probabilities for each of the K baskets, i.e. $\boldsymbol{p}_m = (p_{m1}, \ldots, p_{mK})$ for all $m = 1, \ldots, M$. These scenarios are used in each of the simulation runs for the calibration alongside the basket sample sizes, $\boldsymbol{n_m} = (n_{m1}, \ldots, n_{mK})$ in order to generate data $\boldsymbol{X}$ from $\boldsymbol{X} \sim F(\boldsymbol{p_m}, \boldsymbol{n_m})$.

Each of these simulation scenarios may be weighted to reflect importance or likelihood of them actually occurring in the trial. Thus define weights $\omega_m \in \mathbb{N}$ for each scenario $m = 1, \ldots, M$. If no weight is required, set $\omega_m = 1$ for all $m = 1, \ldots, M$. These weights are integer values, with larger values increasing the number of posterior probabilities that contribute to the calibration process for that scenario, thus increasing the scenarios importance in the calibration relative to other scenarios. If required, these

weights can be normalised to sum to one, i.e. $\omega_m/(\sum_{i=1}^{M}\omega_i)$, however, integer values must be used in Algorithm 1.

Algorithm 1 RCaP - Calibrate Δ_k across several simulation scenarios for any metric, Q.

Data: Total number of simulation scenarios, M, scenarios $\boldsymbol{p_1}, \ldots, \boldsymbol{p_M}$, basket sample sizes $\boldsymbol{n}$, number of simulation runs for each scenario, R, and integer weights for the scenarios, $\omega_1, \ldots, \omega_M$;
Initialisation: $\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_K}$ empty vectors for storing Q
for $m = 1$ to M **do**
  for $r = 1$ to R **do**
    Generate data $\boldsymbol{X} \sim F(\boldsymbol{p_m}, \boldsymbol{n})$
    Fit information borrowing model to obtain posterior densities
    Compute a quantity, Q, obtained from the posterior required for the metric of interest
    for $k = 1$ to K **do**
      if Basket k satisfies the basket specific criterion, $T(\cdot)$ **then**
        for $j = 1$ to ω_m **do**
          $\boldsymbol{Q_k} = \boldsymbol{Q_k} \cup Q$
        end for
      end if
    end for
  end for
end for
$\Delta_k = 100(1-\alpha)\%$ quantile of $\boldsymbol{Q}_k$ for each basket k.
return Cut-off values Δ_k for each basket k;

The generalised novel Robust calibration procedure is described in Algorithm 1, which takes into account the calibration of any metric. The algorithm requires the specification of sample sizes and the definition of all M simulation scenarios under consideration, alongside their weights of importance, ω_m for $m = 1, \ldots, M$. For a simulation scenario, $\boldsymbol{p_m}$, a total of R data sets are generated from $F(\boldsymbol{p_m}, \boldsymbol{n})$. A model is then fit to each of these R data sets to obtain posterior densities. From these posteriors, a quantity Q is computed, where Q is required to compute the metric of interest. A binary basket-specific condition, $T(\cdot)$ is introduced which takes value one when satisfied and zero otherwise. Weights ω_m are utilised in the following step: if basket k satisfies $T(\cdot)$, then ω_m copies of Q under each of the $1, \ldots, K$ baskets are

stored in vectors $\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_K}$. All preceding steps are repeated under each of the M simulation scenarios, thus the higher the weight ω_m, the more scenario m contributes to the vectors $\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_K}$. To compute cut-off values, Δ_k, the appropriate quantile is taken within each of the $\boldsymbol{Q_k}$ vectors. As such, Δ_k will be the quantile of the combined quantities across all M scenarios that satisfy the basket-specific criterion (weighted by importance through ω_m), thereby controlling the metric across all scenarios combined.

When the metric of interest is the type I error rate, the quantity computed is $Q = \mathbb{P}(p_{mk} > q_0|X)$. The probability of a type I error can only be computed when a basket is null, thus the basket-specific condition requires that the true response rate p_{mk} is null. If non-null a type I error cannot occur. When calibrating for type I error control, as is the focus in this chapter, it is therefore important to require that in each of the M simulation scenarios, at least one basket has a null response rate to satisfy the basket specific criterion. The full algorithm applied to control the type I error rate is provided in Section 3.5.3 in Appendix 3.5.

When sample sizes are equal across all or a number of baskets, the Δ_k values will also be equal. In this case, define the set of baskets with equal sample size as E. Of baskets in E, select the basket which satisfies $T(\cdot)$ across the greatest number of the M scenarios. Denote this basket as ϵ, then set $\Delta_k = \Delta_\epsilon$ for all $k \in E$. The RCaP will then maximise the number of simulation scenarios that contribute to the calibration.

Under RCaP in order to control for a type I error, one would expect superior error rate control across non-null cases compared to calibration under the global null, as the Δ_k values obtained will likely be more conservative to ensure error control across multiple scenarios. With the increased conservative nature, it becomes more difficult for the posterior probability $\mathbb{P}(p_k > q_0|X)$ to exceed Δ_k and deem the treatment effective. As such, a decrease in power is also likely.

3.3 Simulation Study

3.3.1 General Setting

In order to explore and compare operating characteristics of the proposed approaches for handling the addition of a new basket to an ongoing trial, numerous simulation studies have been conducted. Simulation studies are split into two categories with the first exploring the case in which vectors of response rates per basket are fixed within the simulation to pre-defined values and secondly, a simulation study in which vectors of response rates are randomly generated within simulation runs. Two lines of comparison are then made within these simulation studies: the calibration under the global null approach is compared to the RCaP followed by a comparison between the approaches for adding a basket to an ongoing trial. In both of these simulations, only subset (a) of PL1 and PL2 in which time of addition is known are considered. However, an exploration into the effect of timing of addition is provided in Section B.6 of the Supporting Information, to assess the performance of PL1(b) and PL2(b).

We consider the following trial setting. There are $K_0 = 4$ existing baskets with $K' = 1$ new basket added part-way through the study. Let the null and target response rates be $q_0 = 0.2$ and $q_1 = 0.4$ respectively. For existing baskets, sample sizes were fixed at $n_{k_0} = 24$ for $k_0 = 1, \ldots, 4$. For the new basket, as the timing of addition is for now assumed as known, assume equal accrual rates across all K baskets and set the sample size as $n_{k'} = 14$ for $k' = 5$. These sample sizes are obtained by a Simon two-stage design (Simon, 1989) with a nominal targeted type I error rate and power of 10% and 80% respectively.

Within each simulation study, the percentage of simulated data sets in which the null hypothesis is rejected in each basket (% Reject) is computed. If the true response rate is q_0, then this value is the type I error rate, otherwise it is the power. Further operating characteristics are presented in Supporting Information B which includes the

family wise error rate, mean point estimates and their standard deviations, as well as the percentage of simulated data sets in which the correct conclusion regarding accepting/rejecting the null was made across all K baskets (% All Correct).

All simulations are conducted using the 'rjags' package v 4.13 (Plummer, 2023) within RStudio v 1.1.453 (R Core Team, 2020), with R v 4.1.2. Simulations consist of 10,000 simulation runs for each data scenario and approach considered.

3.3.2 Prior Specification

Throughout the simulations an independent analysis model is specified such that the prior placed on the logit transformation of the response rate p_k follows a Normal distribution: $\theta_k \sim \text{N}(\text{logit}(0.2), 10^2)$ and is therefore centred around the null response rate with a large variance. The same prior is placed on μ in the exchangeability component of the EXNEX model with a Half-Normal$(0, 1)$ prior placed on σ^2. The prior on the NEX component is specified as in Equation (3.2.2) as suggested by Neuenschwander et al. (2016), where ρ_k (a plausible guess for the true response rate, p_k) is set at a response rate considered as a marginally effective response to treatment, lying between the null and target response rate, $\rho_k = 0.3$. The prior probabilities for assignment to the EX/NEX component are fixed at $\pi_k = 0.5$ for all baskets. Full model specifications are provided in Section 3.5.2 in Appendix 3.5.

3.3.3 Description of the Fixed Data Scenarios Simulation Study

The first simulation study considered is one in which true response rates in scenarios are fixed with each basket having either a null response rate ($p_k = 0.2$) or effective response rate ($p_k = 0.4$). The data scenarios considered are presented in Table 3.3.1. Scenario 1 is the global null state under which all baskets are ineffective to treatment, whereas, scenario 4 is the case where all baskets are truly effective to treatment. Scenarios 2 and 3 are cases in which the new basket is ineffective with a varying number of effective

existing baskets. Similarly, scenarios 5 and 6 are cases in which the new basket is effective to treatment, again with a varying number of effective existing baskets. Results under several other simulation scenarios are presented in Section B.1 of Supporting Information B which covers all global and partial null scenarios, as well as, cases where a varying number of baskets have a marginally effective response to treatment.

Table 3.3.1: Simulation study scenarios: Vectors of true response rates used within the simulation study to compare calibration techniques and approaches for adding a basket.

	p_1	p_2	p_3	p_4	p_5
Scenario 1	0.2	0.2	0.2	0.2	0.2
Scenario 2	0.4	0.2	0.2	0.2	0.2
Scenario 3	0.4	0.4	0.4	0.4	0.2
Scenario 4	0.4	0.4	0.4	0.4	0.4
Scenario 5	0.2	0.2	0.2	0.2	0.4
Scenario 6	0.4	0.2	0.2	0.2	0.4

The cut-off values Δ_{k_0} and $\Delta_{k'}$ are calibrated for each approach separately as described in Table 3.2.1. The calibration under the global null approach means that Δ_{k_0} and $\Delta_{k'}$ are calibrated under scenario 1 to achieve 10% type I error rate. Under RCaP, an average 10% type I error rate is achieved across a number of scenarios. When implementing the RCaP procedure, consideration must be taken into which scenarios to include in the calibration. For the IND, PL1(a) and PL2(a) approaches, an additional two scenarios were incorporated into RCaP alongside those presented in Table 3.3.1: scenario 7 with true response rates $p_k = (0.4, 0.4, 0.2, 0.2, 0.2)$ and scenario 8 with $p_k = (0.4, 0.4, 0.4, 0.2, 0.2)$. With these two scenarios implemented in the RCaP alongside scenarios 1-3 in Table 3.3.1, all global and partial null cases are considered given the new basket has a null response rate. An alternative option is to include all global and partial nulls, taking into account the unequal sample sizes. This would involve including scenarios 1-6 from Table 3.3.1 alongside scenarios 7 and 8 with two further partial nulls: scenario 9 with $p_k = (0.4, 0.4, 0.2, 0.2, 0.4)$ and scenario 10 with $p_k = (0.4, 0.4, 0.4, 0.2, 0.4)$. A simulation study is presented in Section B.2 of Supporting

Information B that compares these two options. Results indicated minimal differences in power and error rates and thus calibration across fewer scenarios is preferred due to the lower computational cost. For the simulation study presented in this chapter, all scenarios carry the same importance and thus weights were set as $\omega_m = 1$ for all scenarios, however, included in Section B.3 of Supporting Information B is an exploration of these weights, demonstrating how operating characteristics changed based on their choice.

Note that calibration under the UNPL approach differs from the other three approaches as its calibration only takes into account the $K_0 = 4$ existing baskets, with the new basket being an unplanned addition. Thus calibration will occur given the following four scenarios: $p_k = (0.2, 0.2, 0.2, 0.2)$ corresponding to scenarios 1 and 5, $p_k = (0.4, 0.2, 0.2, 0.2)$ corresponding to scenarios 2 and 6, $p_k = (0.4, 0.4, 0.2, 0.2)$ corresponding to scenarios 7 and 9 and $p_k = (0.4, 0.4, 0.4, 0.2)$ corresponding to scenarios 8 and 10. These scenarios cover all global and partial nulls given $K = 4$ baskets of equal sample size.

Table 3.3.2: Calibrated Δ_{k_0} and $\Delta_{k'}$ values for IND, UNPL, PL1(a) and PL2(a) under the two separate calibration methods: calibration under the global null and the RCaP.

	Calibration under the global null		RCaP	
	Δ_{k_0}	$\Delta_{k'}$	Δ_{k_0}	$\Delta_{k'}$
IND	0.8599	0.8998	0.9030	0.8989
UNPL	0.8599	0.8599	0.9056	0.9056
PL1(a)	0.8566	0.8409	0.9034	0.9021
PL2(a)	0.8599	0.8409	0.9030	0.9021

Due to the unequal sample sizes across the new and existing baskets, for IND, PL1 and PL2, following Algorithm 1, Δ_{k_0} is selected as the 90% quantile of the posterior probabilities in basket 4 across the implemented scenarios in which its true response is q_0, with $\Delta_{k'}$ selected based on all scenarios considered in the RCaP. Cut-off values for calibration approaches are presented in Table 3.3.2.

3.3.4 Results of the Fixed Data Scenarios Simulation Study

A Comparison of Calibration Approach

Taking this fixed data scenario simulation setting in which six fixed response rate data scenarios were considered, comparisons are first drawn between the two approaches for calibrating Δ_{k_0} and $\Delta_{k'}$: the RCaP and calibrating under the global null. For each of the six fixed scenarios presented in Table 3.3.1 and four approaches for the addition of a basket, the relative difference between the observed type I error rate/power and targeted level (10% and 80% respectively) are measured under each calibration approach. These relative differences are presented in Figure 3.3.1.

First consider the global null scenario, scenario 1, presented in Figure 3.3.1. The calibration under the global null approach achieves exactly the nominal 10% type I error rate, whilst the RCaP reduces the error rate up to 42.5% of the nominal level in existing baskets and 47.4% in the new basket. Under scenario 2, RCaP results in an under-powered study, with up to a 7.8% reduction of the nominal 80% level, however, this came with a 20.1% decrease in type I error rate from the targeted value in existing baskets and 29.3% in the new. Whereas, calibrating under the global null inflates the error rate by up to 30.1% and 36.8% in existing and new baskets respectively with a fairly similar power to the RCaP, although slightly over the nominal level.

The most blatant benefit of the RCaP is observed under scenario 3 in which the new basket is the only one with an ineffective response rate. For this basket, when calibrating under the global null, error rates are almost tripled with values inflated by up to 186.1% of the nominal 10% level, compared to just 31.7% under the RCaP. Under both calibrations, the study is slightly over-powered.

In cases where the new basket is effective (scenarios 4-6), both calibration approaches tend to lead to under-powered estimates in the new basket with the exception of scenario 4, where the power in the new baskets is increased up to 8% of the 80% targeted value across the IND, PL1(a) and PL2(a) approaches when calibrating under the global null.

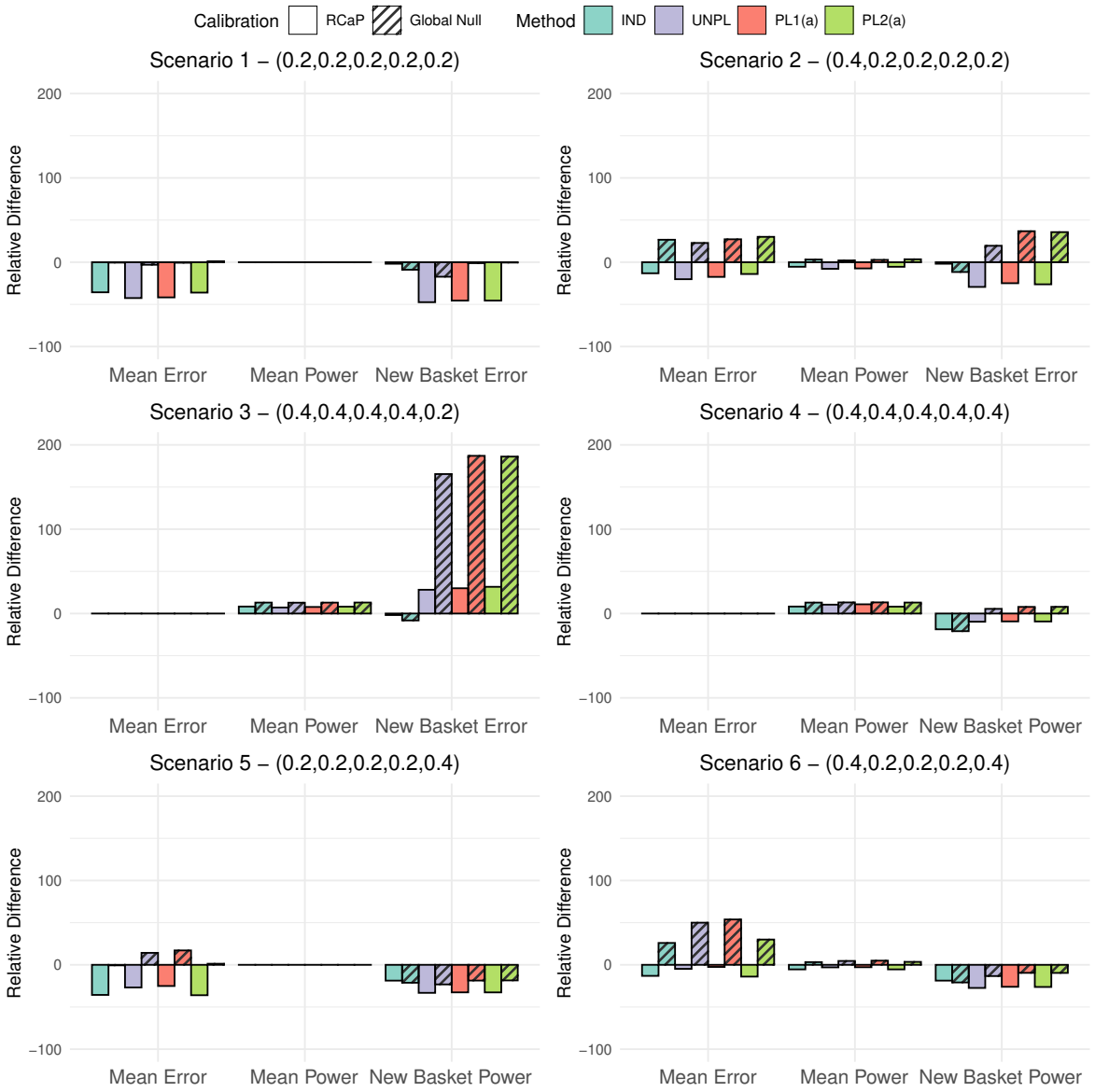


Figure 3.3.1: The relative difference in type I error rate and power compared to the targeted values of 10% and 80% respectively. This is given for all four approaches for adding a basket under the two different calibration schemes, the calibration under the global null and the RCaP. Results are split into 3 categories: mean error in which the percentage of data sets within which the null was rejected is averaged across all ineffective existing baskets; mean power as above but for all effective existing baskets and new basket error/power in which results are the percentage of data sets within which the null was rejected just in the new basket.

For this scenario, RCaP leads to under-powered estimates in the new basket for all four approaches. Power in existing baskets exceeds the nominal 80% value in scenario 4, with slightly higher power observed when calibrating under the global null. Under scenarios

5 and 6, RCaP reduces the type I error rate compared to the nominal level, with a relative difference of up to 36% and 14% reduction in scenarios 5 and 6 respectively. In scenario 6, power in existing baskets is up to a 5.4% reduction of the nominal level using the RCaP compared to an increase of 3.2% under a calibration under the global null approach.

Across the scenarios presented in Figure 3.3.1, estimates in existing baskets are under-powered in two cases (scenarios 2 and 6) with a maximum reduction in power of 7.8% using RCaP. Power in the new basket tends to lie below the nominal 80% level under both calibration approaches. This is due to the smaller sample size of just 14 patients. The new baskets' power is reduced by up to 33.2% under the RCaP compared to 23.2% under the calibration under the global null. However, this comes alongside far superior control of the type I error rate across all baskets on the trial using RCaP. For existing baskets, when calibrating under the global null, the type I error rate has up to a 53.8% increase over the nominal 10% level. Whereas, RCaP controls the type I error rate at or below the nominal level across all considered scenarios for the existing baskets, whilst demonstrating a substantially lower type I error rate in the new basket across all scenarios.

Considering the trade-off observed between error rate control and power improvement under the two calibration approaches, further results presented in this chapter utilises the RCaP to calibrate Δ_{k_0} and $\Delta_{k'}$ in order to provide type I error control. Results for simulation studies in which efficacy criteria are calibrated under the global null are provided in Section B.4 of Supporting Information B.

A Comparison of Approaches for Adding a Basket

Now consider the four approaches for the addition of a basket to an ongoing study under the fixed data scenario setting. The results for power and type I error rate for each approach under the six fixed data scenarios are presented in Figure 3.3.2, which

show the percentage of simulated data sets in which the null hypothesis was rejected. Dashed lines represent both the nominal 10% type I error rate and 80% power. Results for a further ten data scenarios are presented in Section B.1 of Supporting Information B, covering different combinations of effective and ineffective baskets alongside cases in which some baskets have marginally effective response rates.

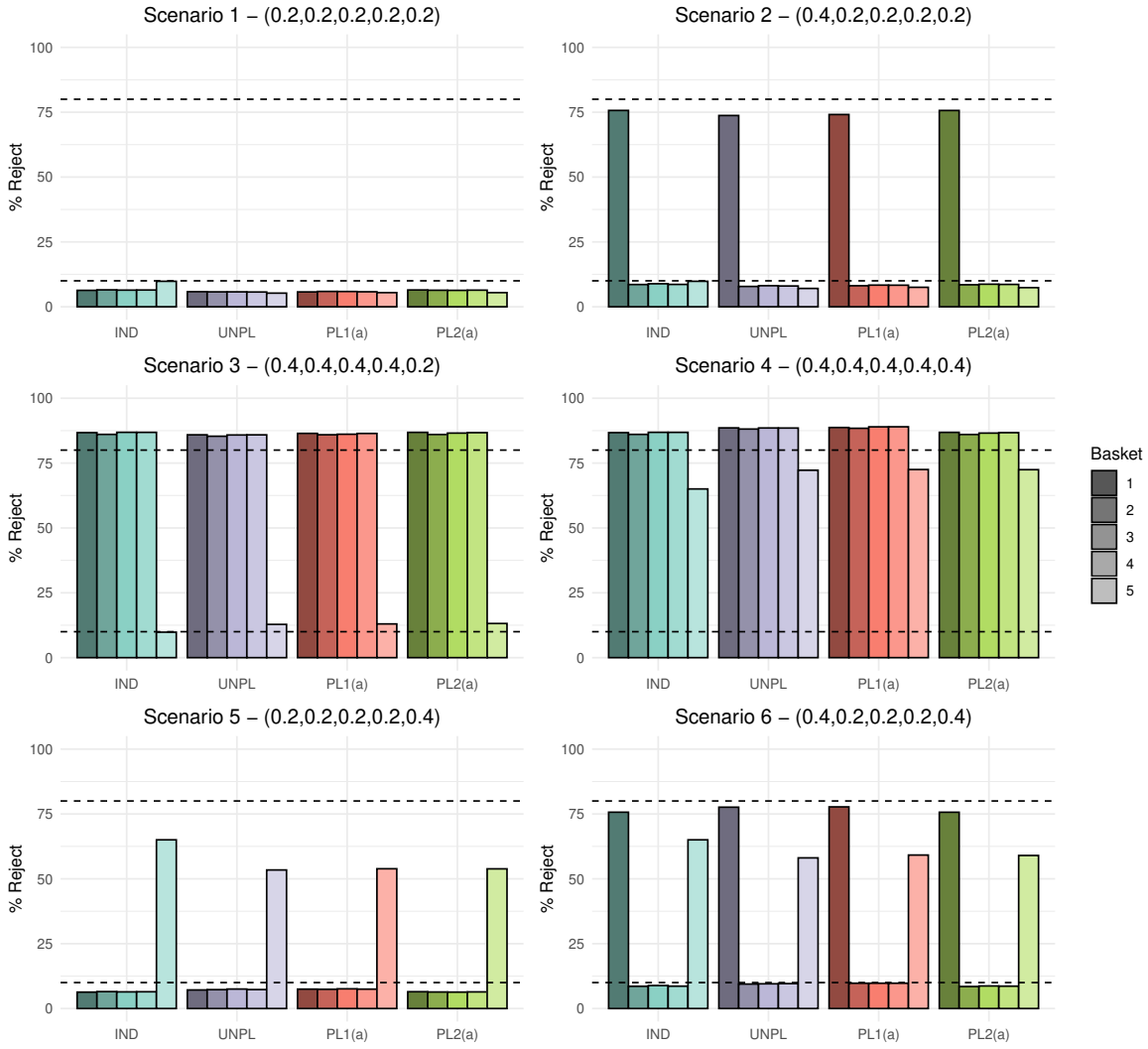


Figure 3.3.2: Fixed scenario simulation study results: The percentage of data sets within which the null hypothesis was rejected, where Δ_{k_0} and $\Delta_{k'}$ were calibrated with RCaP to achieve a 10% type I error rate on average. This is plotted for each of the four approaches for adding a basket in all five baskets.

As Δ_{k_0} and $\Delta_{k'}$ are calibrated using RCaP to achieve an average 10% type I error rate, in some scenarios - including the global null case - the type I error rate lies

below the nominal level. However, under IND, the new basket is always analysed independently and as such, the error rate will remain at the nominal 10% level across all scenarios. Under the global null, for existing baskets, the UNPL and PL1(a) approach in which information is borrowed between all $K = 5$ baskets, results in slightly lower type I error rates in existing baskets compared to other approaches at approximately 5.8%. UNPL, PL1(a) and PL2(a) all have similar error rates in the new basket at around 5.3%.

When analysing existing baskets, IND and PL2(a) are equivalent as they both borrow via the EXNEX model between just the four existing baskets. This is observed in scenario 2 with both approaches giving the highest power at 75.7%, which does lie below the targeted 80% value, but is higher than UNPL and PL1(a) with power of 73.7% and 74.1% respectively. Both UNPL and PL1(a) borrow from the new basket when analysing the existing and hence, as the new basket has a null response rate, the posterior probabilities are pulled down towards the common mean, resulting in lower power. Error rates for all baskets are consistent across approaches with the exception of the IND approach where the new basket type I error is approximately 3% higher.

Scenario 3 shows consistent power in all non-null existing baskets across all four approaches, all above the targeted 80% level. The UNPL approach demonstrates marginally lower power than other methods. The average power under UNPL is 85.7% compared to 86.2% under PL1(a). Both approaches analyse baskets in the same way, borrowing between all K baskets via the EXNEX model, the only difference being the calibration approach. Δ_{k_0} is more conservative under UNPL compared to PL1(a), leading to fewer rejections of the null hypothesis and lower power/error rates. PL1(a) and PL2(a) have marginally higher error rates in the new basket at 13.1% under scenario 3. As $\Delta_{k'}$ is higher under UNPL compared to PL1(a), error is lower at 12.8%.

Under scenario 4, substantial improvements in power is observed in the new basket when information borrowing is utilised. In this scenario, PL1(a) gives the greatest

power for all baskets, with 72.3% power in the new basket. The reduced sample size in basket 5 results in substantially lower power at 65% under an IND approach. A lack of power is also evident for the new basket in scenario 5, however, due the heterogeneity across new and existing baskets, an IND approach gives greatest power at 65%. This is a substantial improvement over the PL1(a) and PL2(a) approaches with power of just 53.8%. Similar findings are present in scenario 6 in terms of the new basket, however both an UNPL and PL1(a) approach give higher power in the existing baskets at 77.7% compared to 75.7% under an IND and PL2(a) analysis.

Overall, the largest difference in power across approaches in all scenarios is just 2%. In the presented scenarios, for existing baskets, the type I error rate is always controlled at or below the nominal level across all approaches. Differences in the type I error rate are observed in the new basket, where the IND approach always controls the type I error rate to the nominal level, whilst error inflation is present under the other three approaches in scenario 3 (type I error rate of around 13%).

3.3.5 Description of the Random Data Scenarios Simulation Study

Based on the results presented in the previous study, no one approach is clearly the most appropriate for use, with little differences observed. Therefore a further simulation study was conducted to distinguish where discrepancies between approaches arise. Within this study, rather than fixing the true response rate for the new basket prior to the trial, it is randomly generated within each trial run of the simulation.

Following the same set-up as the fixed data scenario case, four settings were considered. In each setting the response rates for existing baskets are fixed while the response rate for the new basket is randomly selected with uniform probability across an interval. Three sub-cases are considered in each setting, varying the interval from which p_5 is sampled: sub-case (a) in which the new basket is ineffective to treatment

so $p_5 \in [0.1, 0.2]$, sub-case (b) in which the new basket has an effective response rate so $p_5 \in [0.4, 0.5]$ and finally sub-case (c) in which the new basket is marginally effective to treatment so $p_5 \in [0.2, 0.3]$. The four settings are:

1. Fix the response rate in all the existing baskets as ineffective, i.e. $p_{1,2,3,4} = 0.2$, with p_5 varied across one of the 3 intervals (a), (b) or (c).

2. Fix the response rate in all the existing baskets as effective, i.e. $p_{1,2,3,4} = 0.4$, with p_5 varied across one of the 3 intervals (a), (b) or (c).

3. Fix the response rate in two of the existing baskets as effective, i.e. $p_{1,2} = 0.4$ and two ineffective, i.e. $p_{3,4} = 0.2$, with p_5 varied across one of the three intervals (a), (b) or (c).

4. Fix the response rate in one existing baskets as effective i.e. $p_1 = 0.4$, two as marginally effective i.e. $p_{2,3} = 0.3$ and one as ineffective i.e. $p_4 = 0.2$, with p_5 varied across one of the 3 intervals (a), (b) or (c).

Calibrated Δ_{k_0} and $\Delta_{k'}$ values in Table 3.3.2 are utilised, where calibration is conducted using the RCaP. A total of 12 simulation settings were considered (the four settings outlined above under each of the three sub-cases for sampling p_5) with 10,000 randomly generated data scenarios within each. In each sub-case of the four settings, pair-wise discrepancies between approaches were identified in terms of differing decisions regarding the rejection of the null hypothesis in a basket (and hence differing efficacy conclusions).

3.3.6 Results of the Random Data Scenarios Simulation Study

Results of the 72 pair-wise comparisons across the 12 simulation settings are plotted as several heat maps and presented in Figure 3.3.3. The metric presented is the difference in proportion of correct conclusions made where discrepancies between the two

approaches arose. As an example of a discrepancy, consider setting 1 sub-case (a) in which existing baskets are null, $p_{1,2,3,4} = 0.2$ and the new basket is also null with response rate randomly generated from the interval $p_5 \in [0.1, 0.2]$. A pair-wise discrepancy arises when two approaches give a different conclusion regarding whether or not a basket is effective to treatment, for instance when the IND approach states the treatment is effective in basket five but UNPL states it is ineffective. In this case as the treatment is in-fact ineffective, UNPL led to the correct conclusion thus outperforming IND in this simulation run. Discrepancies are detected across every basket in each of the 10,000 simulation settings and approach which gave the correct inference recorded. Proportions are then taken of correct conclusions in these discrepancies for both approaches under comparison. Each sub-plot within Figure 3.3.3 represents a comparison between two approaches for each of the 12 simulation settings. A negative proportion implies the approach corresponding to the column outperformed the competitor approach in the corresponding row in terms of correct conclusions drawn where discrepancies occur. Within each heat map, the colour of the cell represents the superior approach with brighter colours depicting a greater degree of difference in proportion between the two approaches under comparison.

Consider the pair-wise comparison between IND and UNPL. The IND approach outperforms UNPL by making more correct conclusions in cases of discrepancies in 8 of the 12 simulations. In setting 1 where the existing baskets are null, the difference in approaches is substantial. For example, when the new basket is effective, IND is preferred giving the correct conclusion in 80.9-97% of cases, but when ineffective, UNPL gives correct conclusions in 95.4% of cases where discrepancies lie. Other cases where UNPL is preferred over IND is when there is again homogeneity between existing and new baskets, i.e. in setting 2 where both new and existing baskets are effective. When there is heterogeneity between all baskets, IND in which the new basket is analysed independently tends to outperform the approach that utilises an unplanned addition of

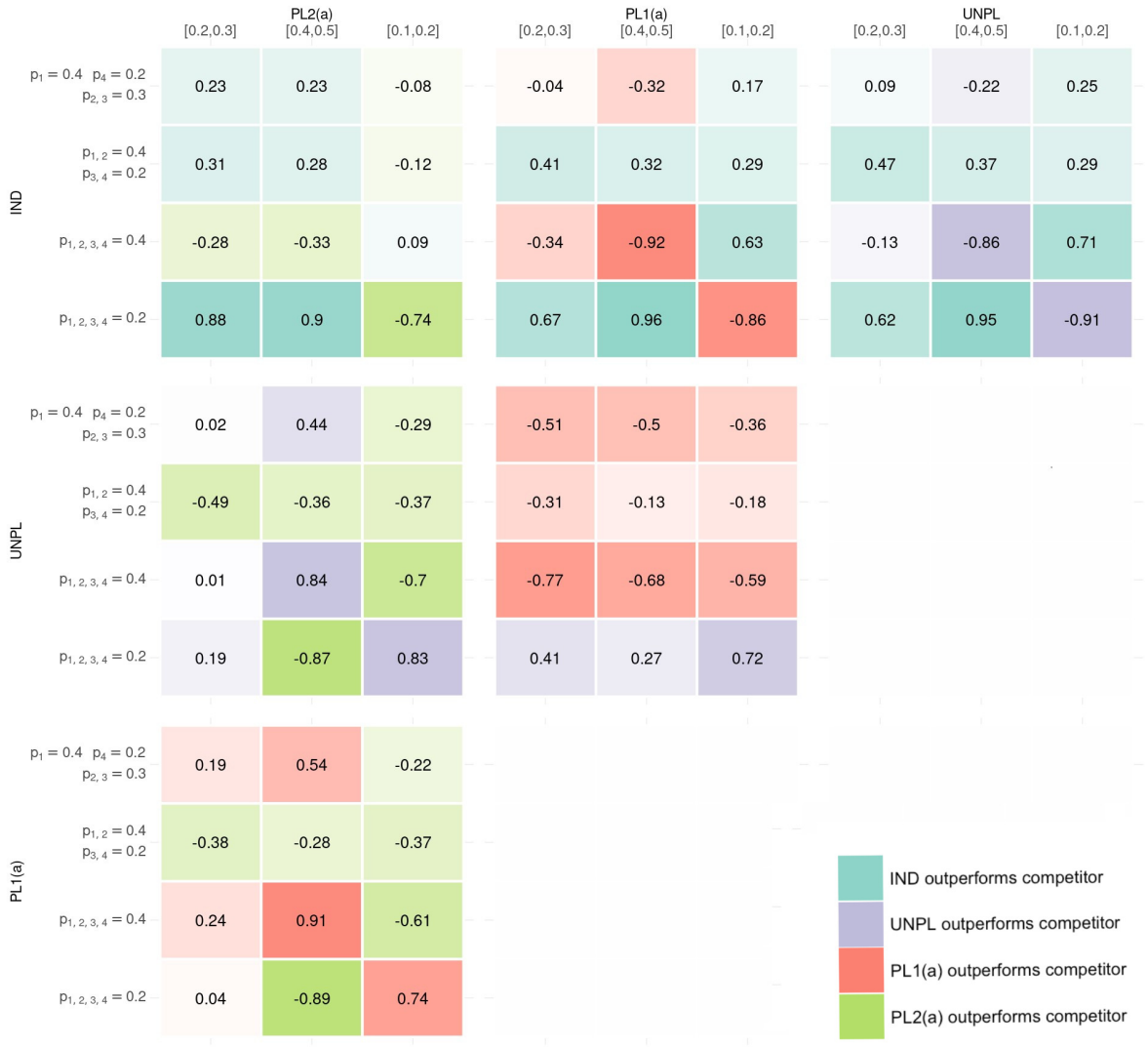


Figure 3.3.3: Pair-wise comparison between approaches in each of the 12 simulation settings within which the true response rate in the new basket is varied. The heat map presents the difference in proportion of times the approach corresponding to rows outperformed the approach corresponding to the column (with negative values indicating the approach in the column gave more correct conclusions over the approach in the row where discrepancies between the two approaches arise). The colour in the heat map represents which approach gave superior correct conclusion, with shade representing the amount of difference between approaches. Blue represents IND giving more correct conclusions where discrepancies lie, Purple for UNPL, Red for PL1(a) and Green for PL2(b).

a basket.

The analysis approach in UNPL is identical to that in PL1(a), the only difference being the calibrated Δ_{k_0} and $\Delta_{k'}$ values. As such, a similar pattern in results to

the IND/UNPL pair-wise comparison are obtained in the comparison between IND and PL1(a). Under UNPL, these cut-off values are more conservative, leading to fewer rejections of the null compared to PL1(a) regardless of whether a basket is truly effective or not. In all cases except setting 1, PL1(a) outperforms UNPL in terms of correct conclusions made. Under setting 1, when the existing baskets are all null, the ideal is for the hypothesis not to be rejected and thus the more conservative Δ_{k_0} value leads to more correct conclusions being made. Breaking down these results further and looking specifically at the new basket only, UNPL leads to correct conclusions in only 3.6-5.4% of discrepancies under setting 1 when the new basket is effective. When ineffective, in all simulation runs, UNPL led to the correct conclusion when discrepancies were identified between the two approaches. So the superior performance in the existing baskets (81.3-85% of correct conclusions) for UNPL in setting 1 overrides the poor performance in the new basket when it is effective. However, in cases where at least one existing basket is effective, PL1(a) gives better correct conclusions over UNPL. This will come from the less conservative cut-off values, leading to more correct rejections and hence higher power.

Under the IND and PL2(a) approaches, analysis for existing baskets is equal and leads to the same conclusions, so the only discrepancies between rejections of the null will occur in the new basket. In settings 2-4 when at least one existing basket is effective, approaches are fairly equal in terms of difference in correct conclusions, with IND performing best when there is heterogeneity between all baskets, with the new basket effective (61.3-65.6% simulation discrepancies where IND gave the correct conclusion over PL2(a)). The Pl2(a) approach has superior performance compared to IND when all baskets are homogeneous.

Similarly, under PL1(a) and PL2(a) analysis for the new basket is equal and thus differences only lie in existing baskets. In cases of complete homogeneity between existing baskets with homogeneity also between the new basket, PL1(a) is the clear

winner as power can be gained through borrowing between all baskets. However, in cases where heterogeneity is observed such as when the new basket is effective and existing ineffective and vice-versa, PL2(a) is superior as it does not draw on information from these heterogeneous baskets when analysing existing baskets. The comparisons between UNPL and PL2(a) result in the same conclusions.

In summary, the IND approach outperforms its competitor approach in 22 out of 36 comparisons, but in particular when heterogeneity is observed. The PL1(a) outperforms IND in the same cases in which PL1(a) also outperforms PL2(a) (i.e. homogeneity between new and existing baskets). In cases where PL2(a) outperforms PL1(a), IND outperformed or performed similarly to PL2(a).

3.4 Discussion

In this chapter, we present four approaches for calibration and analysis of trials when a new basket is added part-way through. Approaches utilise the EXNEX Bayesian information borrowing model which was selected for its flexible borrowing between subsets of baskets.

Through the thorough simulation studies presented, no one of the outlined approaches for adding outperforms its competitors across all cases. An approach which analyses new baskets as independent whilst retaining information borrowing between existing baskets understandably has better error control and power in cases of heterogeneity between new and existing baskets. However, significant power can be gained via information borrowing between all baskets when the new basket is homogeneous to existing ones. This is supported by results from the fixed and random data scenarios. The fixed data scenario simulation results demonstrated that, when the treatment is effective for the population in the new basket, performance of all approaches vary based on the number of effective existing baskets. In our simulation, when at least half of

the existing baskets were effective, higher power was observed in the new basket for the three approaches that implemented information borrowing. However, when very few existing baskets are effective, borrowing information reduces power, thus an independent approach is more appropriate. A key finding was also drawn from the random data scenario simulation study, in that a planned addition of a new basket outperforms an unplanned addition in almost all settings. The exception being the case where all existing baskets are null.

Throughout the simulation studies in this chapter, an assumption is made that the timing of addition of a new basket is known, and thus we assume a fixed sample size in each basket. In practice the calibration of efficacy criteria mostly occurs prior to the trial commencing, and hence before observed sample sizes are available. Due to uncertainty in the observed sample sizes the assumption of fixed sample size has been used to conduct calibration. However, simulation studies in Section B.6 of Supporting Information B explored the case where the timing of addition (and the sample size in the new basket) are unknown. In these simulations, the impact of sample size uncertainty is explored through the monitoring of type I error rate and power as the number of patients in the new basket ranged from 1 up to the sample size of the existing baskets. It is shown that results are fairly robust to the timing of addition, with increased power in new baskets when sample sizes are larger but consistent error and power in existing baskets. This implies the size of the new basket has no detrimental effect on baskets that opened at the commencement of the trial, therefore it is deduced that the main driver of error inflation in the existing baskets, is heterogeneity between the new and existing baskets rather than the sample size. As the sample size increases, the difference in error rates/power between analysing the new basket as independent and conducting information borrowing will decrease, and thus in such a case it may be beneficial to always analyse as independent to avoid issues when heterogeneity arises. In addition, should the impact of much greater or much smaller sample sizes than planned be of

concern, an alternative approach could be to calibrate based on the 'worst case scenario' for the sample sizes (i.e. the sample size which is expected to observe the greatest type I error rate for instance).

Although all simulation studies conducted consisted of a single basket being added alongside four existing baskets, a further simulation study with two existing and two new baskets is presented in Section B.7 of Supporting Information B. Results imply the same conclusions as drawn in the simulation studies presented in this chapter, but with an unplanned addition performing significantly worse than other approaches due to the lack of certainty in the calibration process with only two relatively small baskets being used. It is believed that as the ratio of existing to new baskets increases, the power gained through information borrowing in the new basket further improves due to the gain in certainty around point estimates.

We have also promoted a transition away from the traditional calibration approach in which the type I error rate is controlled under a global null scenario, towards the novel calibration technique, RCaP, presented in this chapter in which the type I error rate is controlled on average across several plausible data scenarios. The concept of calibration across several scenarios is not a wholly new concept and has been implemented extensively in the dose-finding setting for the Continual Reassessment Method (CRM, Lee and Cheung, 2009, 2011). Best et al. (2024), in a similar concept, argued for the use of average type I error rate in the pivotal study setting. Best et al. utilise average type I error rate in order to assess Bayesian designs in which information is borrowed from control or historical data, primarily through informative prior distributions. However, to the best of our knowledge the concept not been implemented in the basket trial setting.

The proposed RCaP provides flexibility by allowing the clinician to specify potential outcomes of the trial in which one would like to control the error rate across, whilst specifying weights to these outcomes to highlight how likely they are to occur and

their importance in the calibration. Throughout the simulation studies presented, equal weights across all scenarios were used. A further exploration of these weights is provided in Section B.3 of Supporting Information B which demonstrates the important role weights play in the RCaP. To summarise the key findings, placing more weight on scenarios with fewer ineffective baskets will produce more conservative cut-off values and with that an improvement in error control but a loss in power, whilst putting more weight on scenarios with mostly ineffective baskets gives less conservative cut-off values and thus higher power.

The advantages of using RCaP over the calibration under the global null approach is not uniform across the scenarios and method of addition implemented. As expected, RCaP is more advantageous over calibrating under just the global null when the scenario differs more substantially from the global null scenario. However, the advantage of superior error control compared to the calibration under the global null approach *is* consistent across all scenarios, with impact on power varied based on the number of effective baskets, showing a small loss in power relative to the targeted value in a handful of cases.

Other adaptive design features such as interim analyses with futility/efficacy stopping are desirable in clinical trials and have been considered across the literature around information borrowing in basket trials including in the work by Jin et al. (2020), Berry et al. (2013), Chu and Yuan (2018) and Psioda et al. (2021). No such design features were included in this chapter. However, the methodology described here could be extended to incorporate such features and future work into this aspect is being conducted. In addition, only a single treatment arm was considered in this work but the methodology can be easily extended to the multi-arm setting in which the treatment is compared to a control group. Similarly, although only a Binomial model is considered within this chapter for modelling response data, more complex models such as an overdispersion model be considered. The goal of implementing such a model would be to control for

heterogeneity within baskets. The impact of using an alternative model hasn't been considered, however, it is believed that the comparison between approaches of addition of a new baskets and comparison between calibration approaches will remain similar, as information borrowing can still be implemented between baskets.

3.5 Appendix

3.5.1 Summary of Approaches for Adding a Basket

Table 3.5.1: Summary of the IND and UNPL approaches for analysis and calibration when adding a basket.

Approach	Description	Calibration	Analysis
IND	Treat the new and existing baskets separately and independent of one another. Analysing the new basket as independent of existing baskets eliminates the potential negative effects of reduced information in the new baskets on existing baskets.	Calibrate Δ_{k_0} based on an EXNEX model applied to the K_0 existing baskets. For new baskets, calibrate $\Delta_{k\prime}$ based on either: (a) independent analysis conducted for each of the $K\prime$ new baskets or (b) borrow information between all $K\prime$ new baskets through a separate EXNEX model.	Analyse in the same way as calibration, with an EXNEX fitted to existing baskets and new baskets analysed with an independent model.
UNPL	Naive approach in which an unplanned addition of new baskets is made and not considered in the calibration procedure. This occurs when it is unknown a basket will be added but it is then believed that borrowing information between all baskets, new and old, can improve inference.	Calibrate Δ_{k_0} based on an EXNEX model applied to the K_0 existing baskets. Fix $\Delta_{k\prime} = \Delta_{k_0}$ once new baskets are added.	Analyse by borrowing information between all K baskets through an EXNEX model.

Table 3.5.2: Summary of the PL1 and PL2 approaches for analysis and calibration when adding a basket.

Approach	Description	Calibration	Analysis
PL1	It is known that a basket will be added during the study and information will be borrowed between all baskets new and existing. This may occur when it is apparent that a basket of patients will benefit from the study, however is not ready in time for the commencement of the trial and thus is planned to be added at a later time.	Calibrate Δ_{k_0} and $\Delta_{k\prime}$ based on an EXNEX model applied to all K baskets. When (a) timing of addition is known: the sample sizes n_k for all K baskets are known and fixed in the EXNEX model. When (b) timing of addition is unknown: further simulation studies are required to explore the effect of $n_{k\prime}$ on operating characteristics, one could calibrate based on the least favourable configuration.	Analyse in the same way as calibration, with an EXNEX model fitted to all K baskets.
PL2	It is known that a basket will be added during the study but when conducting inference on existing baskets only information from other existing baskets is utilised, whereas for inference on new baskets, information is borrowed between all baskets in the trial. This will eliminate the effect of reduced sample sizes in new baskets on estimation of response rates in existing baskets whilst improving power and precision in the new basket.	Calibrate Δ_{k_0} based on an EXNEX model applied to just the K_0 existing baskets. Calibrate $\Delta_{k\prime}$ based on an EXNEX model applied to all K baskets. When (a) the timing of addition is known, sample sizes, n_k, for all baskets are fixed in the calibration procedure. When (b) the timing of addition is unknown, further simulation studies would be required to explore the effect of $n_{k\prime}$ on operating characteristics and adjust calibration accordingly.	Analyse in the same way as calibration with an EXNEX model fitted to just the K_0 existing baskets when analysing existing baskets and with an EXNEX model fitted to all K baskets when analysing the new basket(s).

3.5.2 Model Specification

The trial consists of a total of K baskets, divided into K_0 existing baskets and $K\prime$ new baskets. Parameter choices are those implemented throughout the simulation studies presented in the main text.

IND Calibrate and analyse based on the following model:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K && \delta_{k_0} \sim \text{Bernoulli}(\pi_{k_0}),\\
&\theta_k = \log\Big(\frac{p_k}{1-p_k}\Big), && && M_{1k_0} \sim \text{N}(\mu, \sigma^2),\\
&\theta_{k\prime} \sim \text{N}(-1.39, 10^2), && k\prime = K_0+1, \ldots, K && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_{k_0} = \delta_{k_0} M_{1k_0} + (1-\delta_{k_0}) M_{2k_0}, && k_0 = 1, \ldots, K_0 && \sigma \sim \text{Half-Normal}(0, 1),\\
& && && M_{2k_0} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_{k_0} = 0.5$ for all $k_0 = 1, \ldots, K_0$.

UNPL Calibrate based on the following model:

$$
\begin{aligned}
&Y_{k_0} \sim \text{Binomial}(n_{k_0}, p_{k_0}), && k_0 = 1, \ldots, K_0 && M_{1k_0} \sim \text{N}(\mu, \sigma^2),\\
&\theta_{k_0} = \log\Big(\frac{p_{k_0}}{1-p_{k_0}}\Big), && && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_{k_0} = \delta_{k_0} M_{1k_0} + (1-\delta_{k_0}) M_{2k_0}, && && \sigma \sim \text{Half-Normal}(0, 1),\\
&\delta_{k_0} \sim \text{Bernoulli}(\pi_{k_0}), && && M_{2k_0} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_{k_0} = 0.5$ for all $k_0 = 1, \ldots, K_0$. Analyse based on the following model:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K && M_{1k} \sim \text{N}(\mu, \sigma^2),\\
&\theta_k = \log\Big(\frac{p_k}{1-p_k}\Big), && && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && && \sigma \sim \text{Half-Normal}(0, 1),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && && M_{2k} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_k = 0.5$ for all $k = 1, \ldots, K$.

PL1 Calibrate and analyse based on the following model:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K && M_{1k} \sim \text{N}(\mu, \sigma^2),\\
&\theta_k = \log\Big(\frac{p_k}{1-p_k}\Big), && && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && && \sigma \sim \text{Half-Normal}(0, 1),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && && M_{2k} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_k = 0.5$ for all $k = 1, \ldots, K$.

PL2 Calibrate and analyse existing baskets based on the following model:

$$
\begin{aligned}
&Y_{k_0} \sim \text{Binomial}(n_{k_0}, p_{k_0}), && k_0 = 1, \ldots, K_0 && M_{1k_0} \sim \text{N}(\mu, \sigma^2),\\
&\theta_{k_0} = \log\Big(\frac{p_{k_0}}{1-p_{k_0}}\Big), && && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_{k_0} = \delta_{k_0} M_{1k_0} + (1-\delta_{k_0}) M_{2k_0}, && && \sigma \sim \text{Half-Normal}(0, 1),\\
&\delta_{k_0} \sim \text{Bernoulli}(\pi_{k_0}), && && M_{2k_0} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_{k_0} = 0.5$ for all $k_0 = 1, \ldots, K_0$. Calibrate and analyse new baskets based on the following model:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), && k = 1, \ldots, K && M_{1k} \sim \text{N}(\mu, \sigma^2),\\
&\theta_k = \log\Big(\frac{p_k}{1-p_k}\Big), && && \mu \sim \text{N}(-1.39, 10^2),\\
&\theta_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && && \sigma \sim \text{Half-Normal}(0, 1),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && && M_{2k} \sim \text{N}(-0.85, 4.76),
\end{aligned}
$$

with $\pi_k = 0.5$ for all $k = 1, \ldots, K$.

3.5.3 RCaP: Robust Calibration Procedure for Type I Error Control

Algorithm 2 describes the RCaP specifically for the control of the type I error rate.

Algorithm 2 RCaP - Calibrate Δ_k across several simulation scenarios for type I error rate control

Data: Total number of simulation scenarios, M, scenarios $\boldsymbol{p_1}, \ldots, \boldsymbol{p_M}$, basket sample sizes $\boldsymbol{n_m}$, number of simulation runs for each scenario, R, null response rate, q_0 and integer weights for the scenarios, $\omega_1, \ldots, \omega_M$;
Initialisation: $\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_K}$ empty vectors for storing Q
for $m = 1$ to M **do**
  for $r = 1$ to R **do**
    Generate data $\boldsymbol{X} \sim$ Binomial$(\boldsymbol{p_m}, \boldsymbol{n_m})$
    Fit information borrowing model to obtain posterior densities
    for $k = 1$ to K **do**
      Compute the posterior probability of a type I error $\mathbb{P}(p_{mk} > q_0|X)$, in basket k
      if $T(p_{mk} \leq q_0)$ **then**
        for $j = 1$ to ω_m **do**
          $\boldsymbol{Q_k} = \boldsymbol{Q_k} \cup \mathbb{P}(p_{mk} > q_0|X)$
        end for
      end if
    end for
  end for
end for
$\Delta_k = 100(1-\alpha)\%$ quantile of $\boldsymbol{Q}_k$ for each basket k.
return Cut-off values Δ_k for each basket k;

Chapter 4

Incorporating Historic Information to Further Improve Power When Conducting Bayesian Information Borrowing in Basket Trials

4.1 Introduction

Basket trials have been developed as a form of precision medicine in which an experimental treatment is targeted to a specific genetic make-up rather than a disease type as a whole. This acknowledges that not all patients with the same disease will benefit from a treatment in the same way. This may be due to individual variability in genetics alongside other environmental causes (Ginsburg and Phillips, 2018). Within a basket trial a single treatment is tested on multiple disease types under one master protocol. Each disease type forms a 'basket', with patients across all baskets harbouring the same genetic mutation (Park et al., 2019). Typically such basket trials are implemented in the early stage of the drug development process to assess the efficacy of a treatment on

each of the individual baskets (Tao et al., 2018).

A major advantage of basket trials is the flexibility to test treatments on patients with rare diseases that would not typically warrant their own investigation due to financial and time constraints. However, the small basket sample size that results may cause issues when making inference on treatment effects, particularly in terms of statistical power and precision. Bayesian methodology has been utilised throughout the literature to try and tackle the problem of small sample sizes in basket trials through information borrowing.

Information borrowing utilises an exchangeability concept in that, as all patients in the trial share a common genetic mutation, they will respond homogeneously to the treatment. Prominent methods in the literature implement Bayesian hierarchical models to conduct information borrowing. These methods include the Bayesian hierarchical model (BHM, Berry et al., 2013), calibrated Bayesian hierarchical model (CBHM, Chu and Yuan, 2018), the exchangeability-nonexchangeability model (EXNEX, Neuenschwander et al., 2016) and the modified exchangeability-nonexchangeability model (mEXNEX$_c$, Daniells et al., 2023) to name a few. Empirical Bayesian approaches have also been suggested such as a Bayesian model averaging approach (BMA, Psioda et al., 2021), Fujikawa's design (Fujikawa et al., 2020) and power prior approaches first proposed by Ibrahim and Chen (2000). Such empirical methods have the advantage of analytical posteriors and thus are computationally a lot less intensive.

An alternative approach to improve power and precision is to draw on information from historical or external data sources. Often historical or external information is available for some or all baskets in a trial, where in previous studies the experimental treatment was tested in a similar patient population (Hobbs et al., 2011). An example of such a scenario is the MyPathway study (Hainsworth et al., 2018) which investigated the use Vemurafenib of in BRAFV600 mutation cancers, with the VE-BASKET study (Hyman et al., 2015) also examining the same combination. The two trials had three

baskets (i.e. disease groups) in common. In the wider clinical trial setting, often this historic data is used to inform the control group in a ongoing trial to reduce the number of patients required (Pocock, 1976) or to inform prior distributions (Psioda and Ibrahim, 2019).

Bayesian methods have been used to approach borrowing from historic sources, which typically incorporate historic data into the prior distribution used in the ongoing trial. Most methods down-weight historical data depending on heterogeneity to the current data in the ongoing trial (Bennett et al., 2021). Bennett et al. (2021) and Banbeta et al. (2019) outline a detailed comparison of several methods utilised for the borrowing of historical control data, these include the power prior (PP, Ibrahim and Chen, 2000), modified power prior (MPP, Duan et al., 2006), commensurate prior (Hobbs et al., 2011), robust mixture prior (Schmidli et al., 2014) with the self-adapting mixture prior (SAM Prior, Yang et al., 2023) also an option.

The methods listed above either borrow within a trial or from historic sources but, to the best of our knowledge, none do both simultaneously. It is well known that information borrowing from *any* source can increase the power and precision of treatment effect estimates, thus incorporating both forms of borrowing at once is expected to further benefit power due to the information gained. However, this may come with an inflation in the type I error rate when the assumption of exchangeability between baskets is broken. This occurs when there is heterogeneity between baskets' observed responses. Type I error inflation could also be a result of heterogeneity between current and historic data sources (Kopp-Schneider et al., 2020). To add to this, one must be wary of concerns of bias in historical sources which may arise due to differences in patient populations over time and differing trial conditions (van Rosmalen et al., 2018). We therefore consider it desirable to prioritise and put more weight on borrowing within an ongoing trial than from historic baskets in order to minimise these potential biases.

In this chapter we propose several Bayesian approaches for borrowing between both

current baskets and historic sources under one framework. Note that current baskets refers to baskets that form the ongoing study and historic baskets refers to baskets from historical/external data sources. The proposed approaches include: an EXNEX model where a baskets' probability of exchangeability is determined by the homogeneity between historic baskets; an EXNEX model with a power prior placed on the NEX component; a multi-level mixture model consisting of two EXNEX models (one with historic information and one without); an EXNEX model with pooled historic and current data and a Fujikawa's design which has been adapted to incorporate historic data. Approaches are explored through a simulation study which focuses on binary response data and monitors primarily both basket-wise power and the type I error rate. Simulations are motivated by the MyPathway and VE-BASKET trials presented in Section 4.2. Results display the clear benefit of incorporating the historic information alongside borrowing between current baskets in terms of power gain compared to analysing current data independent of historic data. The results also show a trade-off of this power gain with a slight inflation of error rates, with some approaches demonstrating more inflation than others.

The chapter will be structured as follows: in Section 4.2 we describe a motivating example. In Section 4.3 the novel approaches for historic information borrowing are outlined. A simulation study is presented in Section 4.4.

4.2 Motivating Example

The MyPathway trial (Hainsworth et al., 2018) commenced in 2014 with completion occurring in 2023. This trial consists of multiple non-randomised basket trials under one master protocol. One branch of this trial looked at applying the drug Vemurafenib in patients with solid tumors harbouring the BRAFV600 mutation. Patients with the BRAFV600E-mutated cancers were enrolled across the following baskets: non-small-

cell lung cancer (NSCLC), ovarian cancer, colorectal cancer, anaplastic thyroid cancer and head/neck (larynx) cancer.

Cancer types were classified as either treatment resistant or non-resistant by a steering committee and a Simon's two-stage design (Simon, 1989) for 10% type I error rate and 80% power was used to determine planned sample sizes, with the null and target responses set dependent on the classification of treatment resistance:

- Treatment resistant cancers (e.g. NSCLC): null response rate of 5% and target response rate of 20%, resulting in a sample size of 21 patients per basket.
- Non-treatment resistant cancers (e.g. colorectal or ovarian cancer): null response rate of 10% and target response rate of 25%, resulting in a sample size of 34 patients per basket.

The identical combination of Vemurafenib on patients with BRAFV600 mutation cancers was also studied in the earlier VE-BASKET trial (Hyman et al., 2015) which ran from 2012 to 2014. Both the MyPathway and VE-BASKET trials shared three baskets is common: NSCLC, colorectal cancer and anaplastic thyroid cancer. In the VE-BASKET trial a smaller sample size of 13 patients per basket was planned via a Simon's two-stage design based again on 10% type I error rate and 80% power but with a null and target response rate of 15% and 45% respectively. Observed sample sizes and total responses (both complete and partial) of both the MyPathway and relevant baskets from the VE-BASKET trial are presented in Table 4.2.1.

It appears that both trials were conducted distinctly, with information from the VE-BASKET trial not incorporated into the design or analysis of the MyPathway study. One could argue that the information from the three baskets of common interest could have been utilised in the MyPathway study to inform analysis in some meaningful way, particularly as observed sample sizes were substantially larger. This provides motivation for a trial design that can incorporate borrowing from both current and historic baskets.

Table 4.2.1: Total responses observed (y) and observed sample sizes (n) for baskets in the MyPathway trial and the total responses observed (y^*) and observed sample sizes (n^*) for baskets in the earlier VE-BASKET trial.

Basket	MyPathway *Current*		VE-BAKSET *Historic*	
	y	n	y^*	n^*
NSCLC	6	14	8	20
Colorectal Cancer	1	2	0	10
Anaplastic Thyroid Cancer	1	1	2	7
Ovarian Cancer	2	4	0	0
Head/Neck (Larynx) Cancer)	1	1	0	0

4.3 Methods

4.3.1 Setting

This chapter focuses on non-randomised basket trials with a single treatment arm and binary endpoint, in which a patient either responds to the treatment or does not. Let there be at least one historic basket trial of interest, investigating the same treatment on the same genetic aberration with some baskets in common with the current trial.

Consider a basket trial consisting of K baskets with historic information available for $K^* \in 1, \ldots, K$ of them. For current basket k, there are a total of H_k historic sources of data, where in each past study patients of the same disease type as in basket k received the experimental treatment under investigation. Without loss of generality, assume the first $1, 2, \ldots, K^*$ current baskets have historic information and that current baskets $K^* + 1, \ldots, K$ do not. Responses in a current basket k are denoted by Y_k which follows a Binomial distribution: $Y_k \sim \text{Binomial}(n_k, p_k)$ with sample size n_k and the unknown response rate, p_k, which is the parameter of interest. Similarly, the historic responses also each follow a Binomial distribution. Given that current basket k has historic data from H_k previous studies, denote the basket from historic study j ($j \in \{1, \ldots, H_k\}$) associated with current basket k as $k^{*(j)}$, the responses in basket $k^{*(j)}$ are distributed $Y_{k^{*(j)}} \sim \text{Binomial}(n_{k^{*(j)}}, p_{k^{*(j)}})$ with sample size $n_{k^{*(j)}}$ and response rate $p_{k^{*(j)}}$. Should

only one historic study exist, the superscript (j) is removed and historic data is simply denoted k^* for basket k.

Denote the null response rate in the current trial as q_0. The objective is to test the family of hypotheses:

$$H_0 \ : \ p_k \leq q_0 \quad \text{vs.} \quad H_1 \ : \ p_k > q_0 \qquad k = 1, \ldots, K$$

which is done under a Bayesian framework. In this setting where historic data, D_h, is available, having observed response data D for the current trial, the treatment is deemed effective in basket k if $\mathbb{P}(p_k > q_0 | D, D_h) > \Delta_k$. The decision criteria Δ_k is typically determined through calibration in order to control some metric to a nominal level, which is often the basket-wise type I error rate.

4.3.2 Exchangeability-Nonexchangeability (EXNEX) Model

An approach for information borrowing between baskets on a current trial is the exchangeability-nonexchangeability (EXNEX, Neuenschwander et al., 2016) model. This model provides flexible borrowing between a subset of baskets on the trial, thus not requiring a full exchangeability assumption, allowing for some heterogeneity between baskets. This model does not take into account any historic or external data and considers current baskets only.

The EXNEX model consists of a mixture of two components:

1. Exchangeable (EX) component: Baskets are considered exchangeable within this component and therefore, information borrowing is conducted between them using a Bayesian hierarchical model (BHM, Berry et al., 2013). Basket k is assigned to the EX component with prior probability π_k.

2. Nonexchangeable (NEX) component: Baskets are analysed independently in this component and are considered nonexchangeable with other baskets on the trial.

As such, basket specific priors are placed on the response rate and no information drawn from the other baskets. Baskets are assigned to the NEX component with prior probability $1 - \pi_k$.

The model is presented below, with the log-odds of the response rates modelled as a mixture distribution consisting of the EX and NEX components.

$$Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1, \ldots, K \tag{4.3.1}$$

$$p_k = \delta_k M_{1k} + (1 - \delta_k) M_{2k}, \tag{4.3.2}$$

$$\delta_k \sim \text{Bernoulli}(\pi_k), \tag{4.3.3}$$

$$\theta_{1k} = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)} \tag{4.3.4}$$

$$\mu \sim \text{N}(m_\mu, \nu_\mu), \tag{4.3.5}$$

$$\sigma \sim g(\cdot), \tag{4.3.6}$$

$$\theta_{2k} = \text{logit}(M_{2k}) \sim \text{N}(m_k, \nu_k). \quad \text{(NEX)} \tag{4.3.7}$$

As the EX component is a BHM, response rate estimates within this component are shrunk towards the common mean, μ, with the degree of shrinkage controlled by σ^2. As σ^2 tends to 0, borrowing becomes akin to complete pooling of results, however, as it tends to infinity, stratified analysis of each basket is conducted. Typically it is suggested that a slightly informative prior is placed on μ, for instance by setting m_μ in (4.3.5) to $\text{logit}(q_0)$ with a large variance ν_μ. Several arguments have been made around the choice of prior on σ, with a Half-Normal, Inverse-Gamma or Half-Cauchy density among those suggested. Gelman (2006) argued that the original suggestion of an Inverse-Gamma prior by Berry et al. (2013) had poor behaviour when σ^2 is too close to 0, thus suggested a Half-Cauchy prior instead. Values for the m_k and ν_k parameters in (4.3.7) were suggested by Neuenschwander et al. (2016) as:

$$m_k = \log\left(\frac{\rho_k}{1 - \rho_k}\right), \quad \nu_k = \frac{1}{\rho_k} + \frac{1}{1 - \rho_k}, \tag{4.3.8}$$

where ρ_k is a plausible guess for p_k.

The prior mixture weights, π_k, are often set a priori at $\pi_k = 0.5$ for all k baskets as, ignoring historic information, little to no prior knowledge of the probability of exchangeability is known. Alternatively a Dirichlet prior could be placed on π_k but as stated by Neuenschwander et al. (2016) this has little to no effect on operating characteristics.

4.3.3 EXNEX with a Power Prior in the NEX Component (EXppNEX)

Considering the EXNEX model, homogeneous baskets are assigned to the EX component and information is borrowed between them using a hierarchical model. Power improvement is expected in these baskets due to this borrowing, however, baskets assigned to the NEX component are analysed independently and thus still suffer from the lack of statistical power and precision previously discussed due to their limited sample size. Therefore, it is likely that baskets in the NEX component will benefit more substantially from borrowing from historical data than those already implementing information borrowing in the EX component.

To incorporate historical information into this NEX component when it is available, as opposed to the independent uninformative normal prior being placed on the NEX mixture component (as in (4.3.7)), a power prior can be placed directly on the response rate itself.

A power prior (PP) was first introduced by Ibrahim and Chen (2000) in order to incorporate historical information into a current trial. This is achieved by raising the likelihood of the historical data for each of the $j = 1, \ldots, H_k$ studies to a fixed power, α_j. The power prior for basket k, having observed historic response data $y_{k^{*(j)}}$ for each

of the $j = 1, \ldots, H_k$ historic studies, has the following form:

$$\pi(p_k|\boldsymbol{y}_{\boldsymbol{k}^*}, \boldsymbol{\alpha}) \propto \prod_{j=1}^{H_k} L(p_k|y_{k^{*(j)}})^{\alpha_j} \times \pi_0(p_k), \tag{4.3.9}$$

where $\pi_0(p_k)$ is an initial vague prior on p_k, defined before looking at any historic data and $\boldsymbol{y}_{\boldsymbol{k}^*}$ is the set of historic responses for basket k, whilst $\boldsymbol{\alpha}$ is the set of α_j power values associated with studies $j = 1, \ldots, H_k$. The power values, α_j, are typically bound between 0 and 1 and reflects the expected homogeneity between historic and current data. These α_j parameters are trial specific, allowing some historical studies to carry more weight than others. The selection of a α_j value closer to 0 will move borrowing towards an independent analysis, whilst values close to 1 induce full borrowing. Given the form of (4.3.9), the power parameter, α_j, controls the amount of borrowing as it weights the contribution of the historic data in the posterior parameters (Baumann et al., 2023).

To incorporate this PP into the EXNEX model, the NEX component in (4.3.7) is replaced with a prior which is dependent on the presence of historical data: $M_{2k} = \mathbb{I}_k P_{1k} + (1 - \mathbb{I}_k) P_{0k}$, where $\mathbb{I}_k = 1$ if historic data $y_{k^{*(j)}}$ exists for basket k for some $j \geq 1$, and 0 should no historic information be available for basket k. Now P_{1k} takes the form of the PP, and as such, given an initial Beta(a_k, b_k) prior on p_k for current basket k:

$$P_{1k} \sim \text{Beta}\left(a_k + \sum_{j=1}^{H_k} \alpha_j y_{k^{*(j)}}, b_k + \sum_{j=1}^{H_k} \alpha_j (n_{k^{*(j)}} - y_{k^{*(j)}})\right). \tag{4.3.10}$$

The use of the PP in this way incorporates the historic information, when available, into the model but does not induce borrowing between other baskets on the current trial within the PP itself. To allow for unavailable historic information, P_{0k} is an

uninformative normal distribution placed on the logit-transformed parameter:

$$\theta_{2k} = \text{logit}(P_{0k}) \sim N(m_k, \nu_k),$$

as in (4.3.7) of the EXNEX model. The full model specification is presented in Appendix 4.6.1.

4.3.4 A Multi-Level Mixture Model (MLMixture)

The proposed EXppNEX approach presented in Section 4.3.3 only incorporates historic information in the nonexchangeability component of the EXNEX model and thus baskets assigned to the exchangeable component do not benefit in any way from the historic data. Should all baskets be homogeneous and exchangeable, this historic information is completely disregarded, therefore any potential power gain is wasted. This motivates the need to also incorporate historical information into the EX component to some degree.

One could argue for including historic baskets as distinct baskets in the current trial when conducting analysis, treating them identically to baskets in the ongoing study. When applying the EXNEX model to such a scenario, the historic baskets could be included in the EX's Bayesian hierarchical model, thus inducing borrowing directly from the historic information. However, this ignores the fact that historic baskets correspond to specific baskets in the current trial, inducing the same level of borrowing between a basket and its own historic information as it does between this historic basket and other non-corresponding baskets on the trial. It also puts equal importance of borrowing from historic and current baskets. On the other hand, due to the exchangeability assumption it is assumed a priori that *all* baskets are exchangeable due to the shared genetic component, thus a basket borrowing from its own historic information should be just as acceptable as borrowing from another baskets' historic

data. The mixture weights, π_k, within the EXNEX model should update to assign any heterogeneous historic information into the NEX component in order to restrict borrowing and limit error inflation. However, it is known that the EXNEX model is not sensitive enough to the presence of heterogeneity and thus weights are set too high in this case, inducing too much borrowing resulting in error inflation. This approach would be seen as a more 'extreme' method for borrowing from historic information, which in cases of homogeneity will give substantial improvements in power, but as stated, will likely observe unacceptable error inflation in cases of heterogeneity.

We take this concept of an EXNEX model consisting of all current and historic information and extend it to better handle cases of heterogeneity between current and historic data sources. This is achieved by taking a mixture of such an EXNEX model with a standard EXNEX model that disregards historic information. The mixture weights between these two models will reflect the degree of conflict between the current and historic data. In cases of homogeneity between a current basket and the historic information, mixture weights will shift and put a higher weight on the EXNEX model consisting of historic data, and in cases of heterogeneity, put more weight on the standard EXNEX model, which disregards the heterogeneous historic data. The mixture weights can also be adjusted to put a heavier emphasis on borrowing between baskets on the current trial.

First, to re-emphasise, all baskets current and historic are modelled in the MLMixture model:

$$Y_i \sim \text{Binomial}(n_i, p_i) \qquad i = 1, \ldots, K, 1^{*(1)}, \ldots, 1^{*(H_1)}, \ldots, K^{*(1)}, \ldots, K^{*(H_{K^*})},$$

however, interest lies only in the estimation of the response rates in the current baskets $1, \ldots, K$. Note the subscript has been altered to i as opposed to k in order to distinguish that all current and historic baskets are modelled within this MLMixture model, as historic baskets are treated akin to current baskets in the ongoing trial. We also define

an indicator

$$\psi_i = \begin{cases} 1 & \text{if basket } i \text{ is a historic basket,} \\ 0 & \text{otherwise.} \end{cases}$$

which specifies whether a basket is historic or current.

As stated, the MLMixture model comprises of two EXNEX models, the first of which is denoted $\text{EXNEX}_{\text{all},i}$ which models all current and historic baskets through an EXNEX model, treating historic in the same way as current. The EX component, $\text{EX}_{\text{all},i}$, therefore will consist of a subset of current and historic baskets within which information is shared through a Bayesian hierarchical model. In $\text{EXNEX}_{\text{all},i}$, the NEX component, $\text{NEX}_{\text{all},i}$ is an informative prior based on the observed historic data. If basket i is historic and therefore $\psi_i = 1$, this prior is just an uninformative $\text{Beta}(a_i, b_i)$ prior. As such, the $\text{EXNEX}_{\text{all},i}$ component has the following form:

$$\begin{aligned}
&\gamma_{\text{all},i} = \epsilon_{\text{all},i}\text{EX}_{\text{all},i} + (1 - \epsilon_{\text{all},i})\text{NEX}_{\text{all},i}, \\
&\epsilon_{\text{all},i} \sim \text{Bernoulli}(\pi_{\text{all},i}), \\
&\text{EX}_{\text{all},i} \sim \text{N}(\mu_{\text{all}}, \sigma^2_{\text{all}}), \\
&\mu_{\text{all}} \sim \text{N}(m_{\mu_{\text{all}}}, \nu_{\mu_{\text{all}}}), \\
&\sigma_{\text{all}} \sim g(\cdot) \\
&N_{\text{all},i} \sim \text{Beta}\left(a_i + (1 - \psi_i)\sum_{t=1}^{H_i} y_{i^*(t)}, b_i + (1 - \psi_i)\sum_{t=1}^{H_i}(n_{i^*(t)} - y_{i^*(t)})\right), \\
&\text{NEX}_{\text{all},i} = \text{logit}(N_{\text{all},i}), \\
&\text{EXNEX}_{\text{all},i} = \exp(\gamma_{\text{all},i})/(1 + \exp(\gamma_{\text{all},i})),
\end{aligned}$$

where the mixture weights, $\epsilon_{\text{all},i}$, are updated by the data to reflect the degree of homogeneity between the current and historic baskets. These mixture weights are sampled from a Bernoulli distribution, with the posterior mean close to 1 when basket i is homogeneous to other current and historic baskets, thereby increasing the degree of borrowing

by placing a greater weight on the exchangeability component. As $\pi_{\text{all},i}$ move towards 0, more weight is placed on the nonexchangeability component, which borrows from the historic information but does not borrow information from current baskets. In this first EXNEX model, the historic information appears twice: once in the EX component and once in the NEX component. However, as the weights are binary, each basket is assigned to either the EX or NEX component, therefore the historic information is used only once for each basket.

The second EXNEX model, denoted $\text{EXNEX}_{\text{curr},i}$, does not induce any borrowing from historic data. This model has a very similar form to the $\text{EXNEX}_{\text{all},i}$ model, however, historic baskets are forced into the nonexchangeability component, $\text{NEX}_{\text{curr},i}$, thereby not allowing these baskets into the hierarchy. The response rates in the historic baskets in the second EXNEX model can be estimated through stratified analysis, however, as interest lies only in estimating the response rate in the current baskets, this is ignored as the current baskets do not depend on the historic information in either the EX or the NEX component. The EX component consists of a subset of the current baskets and the $\text{NEX}_{\text{curr},i}$ component is an uninformative $\text{Beta}(a_i, b_i)$ prior, therefore, also ignores historic data. The $\text{EXNEX}_{\text{curr},i}$ model has the following form:

$$
\begin{aligned}
&\gamma_{\text{curr},i} = \epsilon_{\text{curr},i}\text{EX}_{\text{curr},i} + (1 - \epsilon_{\text{curr},i})\text{NEX}_{\text{curr},i},\\
&\epsilon_{\text{curr},i} \sim \text{Bernoulli}((1 - \psi_i)\pi_{\text{curr},i}),\\
&\text{EX}_{\text{curr},i} \sim \text{N}(\mu_{\text{curr}}, \sigma^2_{\text{curr}}),\\
&\mu_{\text{curr}} \sim \text{N}(m_{\mu_{\text{curr}}}, \nu_{\mu_{\text{curr}}}),\\
&\sigma_{\text{curr}} \sim f(\cdot)\\
&N_{\text{curr},i} \sim \text{Beta}(a_i, b_i),\\
&\text{NEX}_{\text{curr},i} = \text{logit}(N_{\text{curr},i}),\\
&\text{EXNEX}_{\text{curr},i} = \exp(\gamma_{\text{curr},i})/(1 + \exp(\gamma_{\text{curr},i})),
\end{aligned}
$$

where mixture weights, $\epsilon_{\text{curr},i}$ are set to 0 for all historic baskets

$$i = K, 1^{*(1)}, \ldots, 1^{*(H_1)}, \ldots, K^{*(1)}, \ldots, K^{*(H_{K^*})}.$$

For current baskets $1, \ldots, K$, these mixture weights now only reflect the level of homogeneity between itself and all other current baskets.

To fit the MLMixture model, both $\text{EXNEX}_{\text{all},i}$ and $\text{EXNEX}_{\text{curr},i}$ are fit distinctly. The posterior for basket k is then a mixture of the posteriors obtained under both models:

$$\begin{aligned} p_k &= \lambda_k \text{EXNEX}_{\text{all},k} + (1 - \lambda_k)\text{EXNEX}_{\text{curr},k}, \\ \lambda_k &\sim \text{Bernoulli}(\pi_{\lambda,k}), \end{aligned}$$

where λ_k reflects the degree of homogeneity between a current basket k and its own historic baskets' $k^{*(j)}$, as well as, the homogeneity to other baskets' historic data. Values of $\pi_{\lambda,k}$ close to 1 can induce a higher level of borrowing from historic baskets, whilst values close to 0 analyse current baskets as independent from any historic data. This λ_k value will not measure the degree of homogeneity between current baskets as there is potential to borrow between these baskets in both sides of the mixture.

This model form provides flexibility, allowing baskets with historic sources to borrow between both current baskets and all historic data, whilst letting baskets without historic information to also gain from the historic information of other exchangeable baskets. Similarly, should data be heterogeneous, the model has the option of analysing as completely independent. A downside of this approach is its computational intensity as the extra layers of mixture and increased number of variables increases the model complexity. The models were fit using the 'rjags' package v 4.12, (Plummer, 2023) within RStudio v 1.1.453 (R Core Team, 2020) and required each basket to be modelled separately, thus computation time will further grow as both K and K^* increase.

A further discussion on computation time and a comparison between approaches is provided in the Section C.3 of Supporting Information C.

4.4 Simulation Study

Two approaches for incorporating historic information have been proposed and outlined in Section 4.3, in this section we aim to explore the operating characteristics of each model to compare performance. Performance of approaches are assessed using extensive simulation studies motivated by the MyPathway and VE-BASKET trials as described in Section 4.2. As such, in the simulation study there are a total of $K = 5$ current baskets with historical information for the first $K^* = 3$, also assume that $H_k = 1$ for $k = 1, 2, 3$ so that when historic information is available, there is only a single source of historic data, thus any superscripts (j) may be dropped for notation sake. Sample sizes are fixed and equal across the current baskets $k = 1, \ldots, K$ at $n_k = 34$, with a null and target response rate of $q_0 = 0.1$ and $q_1 = 0.25$ respectively. Sample sizes for each of the historic baskets are $n_{k^*} = 13$. As in the MyPathway study, the target/nominal type I error rate and power are 10% and 80% respectively.

Within the simulation study, responses are randomly sampled based on a true response rate, whilst historic data is fixed. This is done to mimic a trial setting where simulation studies are conducted prior the current trial, at which time the historic information has already been observed. A total of 8 true response rate data scenarios were considered for the current data and are presented in Table 4.4.1. Scenario 1 represents the global null in which all baskets are ineffective against the treatment, whereas, scenario 6 is the global alternative under which all are effective. Scenarios 2-5 cover partial nulls in which an increasing number of baskets are effective to treatment. Scenarios 7 and 8 both consider cases where one of the baskets without historic information is effective and varied the effectiveness in baskets with historic information.

Table 4.4.1: True response rate data scenarios considered in the simulation study for comparison of novel approaches to historic information borrowing.

Scenario	p_1	p_2	p_3	p_4	p_5
1	0.10	0.10	0.10	0.10	0.10
2	0.25	0.10	0.10	0.10	0.10
3	0.25	0.25	0.10	0.10	0.10
4	0.25	0.25	0.25	0.10	0.10
5	0.25	0.25	0.25	0.25	0.10
6	0.25	0.25	0.25	0.25	0.25
7	0.10	0.10	0.10	0.25	0.10
8	0.25	0.10	0.10	0.25	0.10

Each of these 8 data scenarios are split into four sub-cases consisting of four different historic data settings, resulting in a total of 32 simulation scenarios. Each setting differs the number of effective historic baskets from 0 up to all three effective and are presented in Table 4.4.2

Table 4.4.2: Historic data settings considered in the simulation study for comparison of novel approaches to historic information borrowing.

Sub-case	y_{1^*}	y_{2^*}	y_{3^*}
(a)	1	1	1
(b)	3	1	1
(c)	3	3	1
(d)	3	3	3

Efficacy is determined using posterior distributions, so having observed current data D and historic data D_h, basket k is deemed sensitive to the treatment if $\mathbb{P}(p_k \geq 0.1|D, D_h) \geq \Delta_k$. Traditionally this efficacy cut-off Δ_k would be calibrated under the global null scenario in which all baskets have a true null response rate in order to control the basket-wise type I error. However, this simulation study implements the Robust Calibration Procedure (RCaP) as outlined in Chapter 3, in order to achieve an average basket-wise type I error rate of 10% across a number of scenarios as opposed to under just the null. RCaP is taken across all 8 scenarios presented in Table 4.4.1 to produce cut-off values Δ_k. Note that Δ_k values are calibrated separately for each of

the four sub-scenarios, i.e. for each of the four observed historical data settings and for each of the six approaches. The calibrated Δ_k values are presented in Table 4.6.1 in Appendix 4.6.2, and a description of the RCaP procedure implemented in this study is provided in Section C.1 of Supporting Information C.

For all 32 scenarios considered, the following operating characteristics were computed:

- % Reject: the percentage of simulated data sets in which the null hypothesis was rejected. If the true response rate is null, this is the basket-wise type I error rate, else it is the power.
- Family-wise error rate (FWER): the percentage of simulated data sets in which at least one truly null basket was deemed sensitive to treatment.
- % All correct: the percentage of simulated data sets in which the correct efficacy conclusion was made across all K baskets.
- Mean of the response rate point estimate across all simulated data sets. The standard deviation of these point estimates across data sets is also provided.

Results in the main text focus on the type I error and power, i.e the % Reject values, however, results of the other metrics listed above are provided in Supporting Information C.

A total of 5,000 simulations were run for each of the 32 scenarios using the 'rjags' package v 4.12, (Plummer, 2023) within RStudio v 1.1.453 (R Core Team, 2020). For each simulation run, for methods that utilise an MCMC approach, the MCMC was conducted with 100,000 iterations.

4.4.1 Adapted Fujikawa's Design (histFujikawa)

For comparison within the simulation study, an alternative empirical approach is also considered. Fujikawa et al. (2020) developed a closed-form Beta posterior for each

basket, within which the parameters incorporate the response data of other baskets in the trial in order to borrow information. Fujikawa's design first fits independent Beta-Binomial models to the response rates, p_k, with a Beta(a_k, b_k) prior used. This results in the following posterior: $\pi(p_k|Y_k = y_k) = \text{Beta}(a_k + y_k, b_k + n_k - y_k)$ for basket k. Fujikawa proposed borrowing of information by taking the weighted sum of these posterior parameters with weight, $\omega_{k,i}$ representing the degree of homogeneity between baskets k and i. The posterior of the response rate given observed data D for all baskets is then

$$\pi(p_k|D) = \text{Beta}\left(\sum_{i=1}^{K} \mathbb{I}(\omega_{k,i}^{\epsilon} > \tau)\omega_{k,i}^{\epsilon}(a_k + y_i), \sum_{i=1}^{K} \mathbb{I}(\omega_{k,i}^{\epsilon} > \tau)\omega_{k,i}^{\epsilon}(b_k + n_i - y_i)\right), \tag{4.4.1}$$

where $\epsilon \geq 1$ and $\tau \in [0, 1]$ are tuning parameters. ϵ controls how quickly the weights move to 0 to discourage borrowing as baskets become increasingly heterogeneous to one another, whilst τ acts as a cut-off which sets the weights to 0 should heterogeneity cross the threshold, τ.

Fujikawa suggested setting the weights $\omega_{k,i}$ based on 1 minus the pair-wise Jensen-Shannon divergence (JSD) between the Beta-Binomial posteriors for baskets k and i:

$$\omega_{k,i} = 1 - \text{JSD}(\pi(p_k|Y_k = y_k), \pi(p_i|Y_i = y_i)), \tag{4.4.2}$$

where the JSD between two distributions P and Q is

$$\text{JSD}(P, Q) = 1/2(KL(P||M) + DL(Q||M))$$

with $M = 1/2(P + Q)$ (Fuglede and Topsoe, 2004).

$$KL(P||M) = \sum_{x} P(x) \log(P(x)/M(x))$$

is the Kullback-Leibler Divergence (KLD). In order to obtain weights bounded between 0 and 1, KLD is used with base 2 logarithm (Baumann et al., 2023).

A similar approach can be used to also incorporate historical information, with the parameters of the Beta posterior now including both a weighted sum of current data and a weighted sum of historic data, with weights determined as a function of homogeneity between baskets:

$$\pi(p_k|D, D_h) = \text{Beta}\Bigg(a_k + \sum_{i=1}^{K}\left(\mathbb{I}\{\omega^{\epsilon}_{k,i} > \tau\}\omega^{\epsilon}_{k,i}y_i + \zeta_k \sum_{j=1}^{H_k}\left(\mathbb{I}\{\omega^{\epsilon}_{k,i^*(j)} > \tau\}\omega^{*\epsilon}_{k,i^*(j)}y_{i^*(j)}\right)\right),$$
$$b_k + \sum_{i=1}^{K}\left(\mathbb{I}\{\omega^{\epsilon}_{k,i} > \tau\}\omega^{\epsilon}_{k,i}(n_i - y_i) + \zeta_k \sum_{j=1}^{H_k}\left(\mathbb{I}\{\omega^{*\epsilon}_{k,i^*(j)} > \tau\}\omega^{*\epsilon}_{k,i^*(j)}\left(n_{i^*(j)} - y_{i^*(j)}\right)\right)\right)\Bigg), \tag{4.4.3}$$

where weights between current data sources, $\omega_{k,i}$ are computed as in (4.4.2) using the JSD. Weights between a current basket k and a historic basket associated with basket i from one of the H_i previous studies are also computed using JSD but are set to 0 should no historic information be available for basket i:

$$\omega_{k,i^*(j)} = \begin{cases} 1 - JSD\left(\pi(p_k|Y_k = y_k), \pi(p_{i^*(j)}|Y_{i^*(j)} = y_{i^*(j)})\right) & \text{If historic information,} \\ & \text{is available for basket } i \\ 0 & \text{Otherwise.} \end{cases}$$

Note that unlike in Fujikawa's design, information is not shared between the prior distributions, with the prior parameters moved outside of the sum. The tuning parameters ϵ and τ are still defined in the same way as in Fujikawa's design, with ϵ defining the degree of decline of weights with heterogeneity and τ as a cut-off for borrowing when the degree of heterogeneity becomes too large. This approach has a computational advantage over the other methods proposed due to the closed form solution of

posteriors it provides.

4.4.2 Other Competing Approaches

An alternative approach to integrate historic information into a borrowing model between current baskets, is one in which the historic data is used to define the probabilities of exchangeability prior to observing data from the current baskets. These prior probabilities, π_k, are then used as mixture weights in (4.3.3) in the EXNEX model to analyse data from the current baskets. Therefore, within this approach, denoted mEXNEX$_{\text{hist}}$, historical information is not borrowed from directly, and data is not used in the analysis model itself beyond updating the mixture weights.

This can be viewed as a version of the modified exchangeability-nonexchangeability (mEXNEX$_c$, Daniells et al., 2023) model, within which a baskets' exchangeability weight is computed using a data-driven approach. The original mEXNEX$_c$ approach first applies simple independent Beta-Binomial models to each basket, then utilises the average Hellinger distance between the resulting posteriors in order to compute π_k. For this approach, rather than computing the Hellinger distance between current baskets, it is computed for posteriors of pooled historic response data. To find such posteriors, for each of the K^* historic baskets, pool the results of all H_k studies associated with basket k and define $\hat{y}_{k^*} = \sum_{j=1}^{H_k} y_{k^*(j)}$ and $\hat{n}_{k^*} = \sum_{j=1}^{H_k} n_{k^*(j)}$. Simple Beta-Binomial models are fit to each of the historic responses $\hat{y}_{k^*}$ with an uninformative Beta(1,1) prior implemented.

For baskets with historic information, the probability of exchangeability is set as the average Hellinger distance between all historic baskets:

$$\pi_k = \sum_{i^*=1, i^* \neq k^*}^{K^*} \frac{1 - h_{i^*,k^*}}{K^* - 1} \quad \text{for } k = 1, \ldots, K^*, \tag{4.4.4}$$

where h_{i^*,k^*} is the Hellinger distance between the historic baskets (consisting of pooled

data) for two current baskets i and k. For baskets without historic information, define their probabilities of exchangeability as:

$$\pi_k = \zeta_k \sum_{i=1}^{K^*} \frac{\pi_i}{K^*} \quad \text{for } k = K^* + 1, \ldots, K, \tag{4.4.5}$$

i.e. as the average probability of exchangeability of those baskets that *do* have historic data available, down-weighted by a scalar, $\zeta_k \in [0, 1]$. The purpose of this ζ_k is to account for uncertainty in the probability of exchangeability in baskets in which there is no previously observed data. Due to the exchangeability assumption, a priori it is believed that those with and without historic data are exchangeable and thus in this approach it is assumed the π_k values will be similar for those without historic data. However, this assumption may not hold, in which case the exchangeability for these baskets without previous data may not equate to those with historic data. The scalar ζ_k limits the potential impact this could have on inflated error rates.

This model is also included in the simulation study alongside histFujikawa for comparison purposes.

4.4.3 Prior and Parameter Choices

The models for comparison are as follows:

1. **EXNEX**: the EXNEX model independent of any historic information as described in Section 4.3.2.

2. **EXNEX$_{\text{pool}}$**: an EXNEX model which incorporates historic information by pooling the results of basket k and all results from the H_k previous studies. An EXNEX model as described in Section 4.3.2 is applied to the pooled responses. This takes into account that the historic baskets are associated with a specific basket on the current trial.

3. **mEXNEX$_{\text{hist}}$**: a modified EXNEX approach with data-driven exchangeability weights based on the historic data as described in Section 4.4.2.

4. **EXppNEX**: an EXNEX model with a power prior placed on the NEX component as in Section 4.3.3.

5. **MLMixture**: a multi-level mixture model consisting of two EXNEX models as in Section 4.3.4.

6. **histFujikawa**: an adapted Fujikawa's design as outlined in Section 4.4.1 in which the closed-form posterior incorporates information from both current and historic data.

All six methods explored in the simulation study have several prior and parameter choices. Table 4.4.3 summarises such choices for all methods. Full model outlines are given in Appendix 4.6.1. For the EXNEX, EXNEX$_{\text{pool}}$, EXppNEX and the MLMixture models equal mixture weights of 0.5 were utilised throughout to fully allow the model to update the weights based on homogeneity/heterogeneity. Other values of π_{λ_k}, $\pi_{\text{all},i}$ and $\pi_{\text{curr},i}$ were considered for the MLMixture model and further discussed in Section 4.4.5, however, the choice of equal weights throughout demonstrated a good balance between error control and power improvement. A weight metric based on JSD as in Fujikawa's design was considered to shift these weights in the EXppNEX and the MLMixture models, based on homogeneity of response data, but this inflated error and decreased power in some cases, thus fixing the weights proved superior, whilst also reducing the model complexity. For EXppNEX the power prior parameter, α, was set at 0.5 in order to discount the historical information in the informative prior. Alternative α values of 0.25 and 1 were also considered and are discussed in Section 4.4.5, the findings of which suggest setting $\alpha = 0.5$ as a reasonable choice.

For all methods bar histFujikawa, a hierarchy is placed on an EX component within which hyper-priors are placed on the common mean μ and borrowing parame-

Table 4.4.3: Prior and parameter choice for the simulation study for comparison of novel approaches to historic information borrowing.

Model	Parameters and Priors
EXNEX	$\pi_k = 0.5$ for $k = 1, \ldots$,K, $m_\mu = \text{logit}(0.1)$, $\nu_\mu = 10^2$, $g(\cdot)$=Half-Normal(0,1), m_k and ν_k are computed as in (4.3.8) with $\rho_k = 0.2$.
$\text{EXNEX}_{\text{pool}}$	$\pi_k = 0.5$ for $k = 1, \ldots$,K, $m_\mu = \text{logit}(0.1)$, $\nu_\mu = 10^2$, $g(\cdot)$=Half-Normal(0,1), m_k and ν_k are computed as in (4.3.8) with $\rho_k = 0.2$.
$\text{mEXNEX}_{\text{hist}}$	$m_\mu = \text{logit}(0.1)$, $\nu_\mu = 10^2$, $g(\cdot)$=Half-Normal(0,1), m_k and ν_k are computed as in (4.3.8) with $\rho_k = 0.2$, ζ_k=0.8.
EXppNEX	$\pi_k = 0.5$ for $k = 1, \ldots$,K, $m_\mu = \text{logit}(0.1)$, $\nu_\mu = 10^2$, $g(\cdot)$=Half-Normal(0,1), m_k and ν_k are computed as in (4.3.8) with $\rho_k = 0.2$, $a_k = b_k = 1$, $\alpha_j = 0.5$ for $j = 1, \ldots, H_k$ for all k.
MLMixture	$\pi_{\lambda,k} = 0.5$ for $k = 1, \ldots, K$, $\pi_{\text{all},i} = \pi_{\text{curr},i} = 0.5$ and $a_i = b_i = 1$ for $i = 1, \ldots, K, 1^*, \ldots, K^*$, $m_{\mu_{\text{all}}} = m_{\mu_{\text{curr}}} = \text{logit}(0.1)$, $\nu_{\mu_{\text{all}}} = \nu_{\mu_{\text{curr}}} = 10^2$, $g(\cdot)$=$f(\cdot)$=Half-Normal(0,1).
histFujikawa	$a_k = b_k = 1$ for $k = 1, \ldots, K$, $\epsilon = 2$, $\tau = 0.2$, $\zeta_k = 0.8$, weights $\omega_{k,i}$ and ω_{k,i^*} are determined using JSD.

ter σ. For each of the approaches which possess an EX component, the hyper-priors $\mu \sim \text{N}(\text{logit}(q_0), 10^2)$ and $\sigma \sim$ Half-Normal$(0, 1)$ are applied. For both histFujikawa and $\text{mEXNEX}_{\text{hist}}$, a historic scalar of $\zeta_k = 0.8$ is implemented to down-weight the contribution of historic data to borrowing in the former and to reduce the prior borrowing probability for baskets with unobserved historic data in the latter. The tuning parameters in histFujikawa are set at $\epsilon = 2$ to provide moderate reduction in weights when posteriors become increasingly dissimilar and τ is set at 0.2.

4.4.4 Simulation Results

The results under four of the eight scenarios are presented in Figures 4.4.1 and 4.4.2, which show the type I error rate and power for each of the five baskets under all six information borrowing approaches. Dashed lines are provided on each plot to highlight

the nominal 10% type error rate and 80% power. A dashed line is also placed on a value of 90% in order to distinguish power improvement between approaches when the nominal level is exceeded. Scenarios 2 and 5 were selected as they demonstrate the more 'extreme' cases wherein just a single baskets is effective or ineffective respectively. These two scenarios tend to give the lowest power and highest error inflation respectively and thus assessment of the performance of the approaches under these two 'extreme' cases is most compelling. Scenario 6 is the global alternative and will best demonstrate power improvement across the approaches. Scenario 8 differs as it consists of a basket without historic information being effective to the treatment alongside just a single effective basket with historic information. Thus this scenario allows a comparison in power dependent on whether or not a basket has historic information. The plotted results of the remaining four scenarios are given in Appendix 4.6.3.

Beginning with scenario 2 (presented in the top half of Figure 4.4.1), the EXNEX model demonstrates sub-par power for the one effective basket, lying below the nominal level at 78.9%. Both the mEXNEX$_{\text{hist}}$ and histFujikawa methods also fail to reach the nominal level in almost all cases with power as little as 66.9% and 70.2% respectively. In cases (a) and (d), responses in historic baskets are completely homogeneous, thus in the mEXNEX$_{\text{hist}}$ approach, the probabilities of exchangeability π_k are set at 1 for baskets with historic information and $\pi_k = 0.8$ in baskets 4 and 5. This results in strong borrowing, thus the posterior for the one and only effective basket is pulled down towards those 4 ineffective baskets, resulting in a loss in power. Similarly, the histFujikawa design lacks power in this case due to level of heterogeneity observed between basket 1 and other current baskets on the trial, causing weights $\omega_{k,i}$ to be at or close to 0, moving analysis to an independent analysis taking the form of an uninformative Beta-Binomial model.

All approaches under scenario 2 have type I errors at or below the nominal 10% level. Only minute differences are observed between the EXNEX$_{\text{pool}}$, EXppNEX, and MLMix-

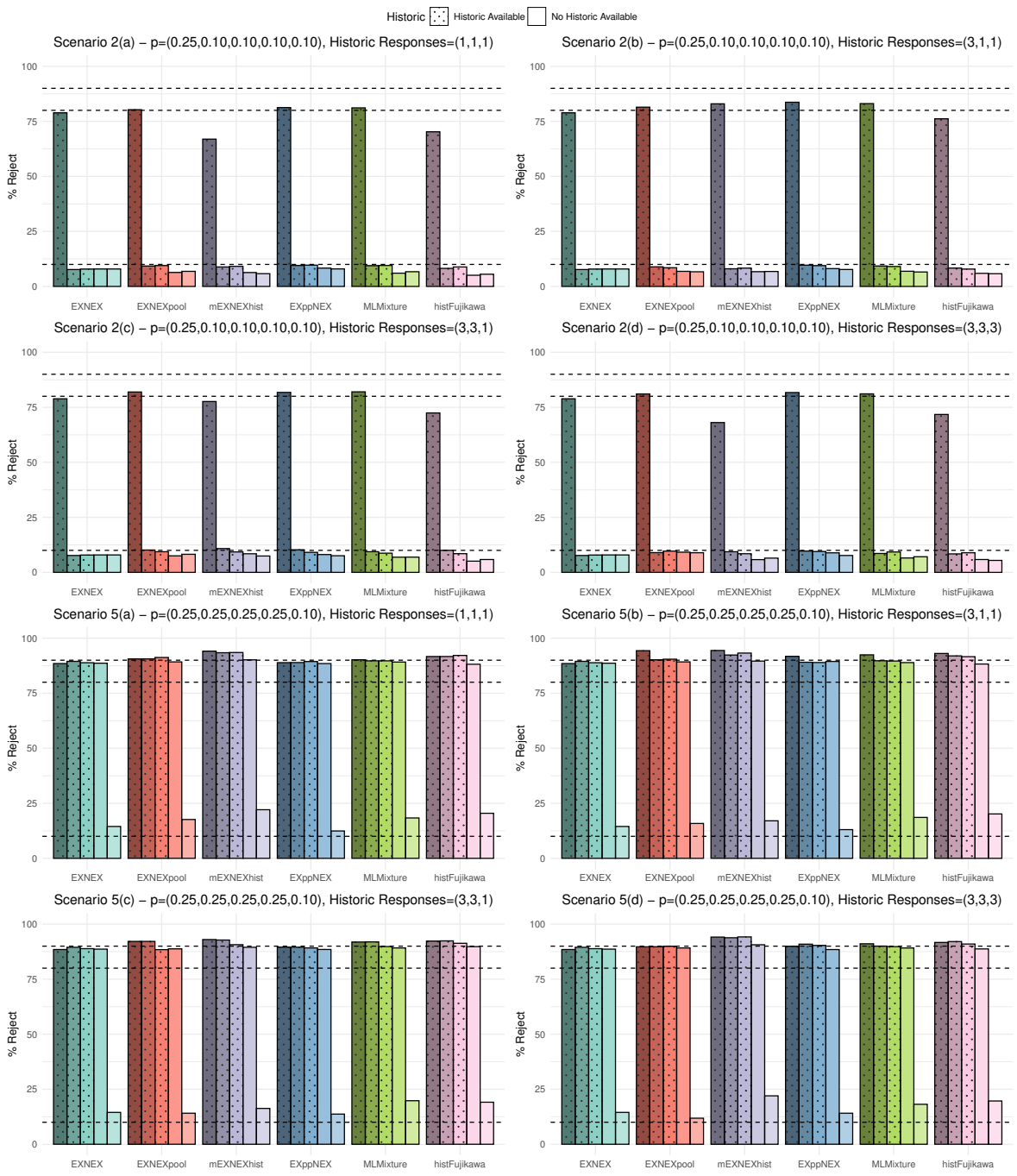


Figure 4.4.1: Type I error rate and power under each of the 6 approaches for historic information borrowing for scenarios 2 and 5 cases (a)-(d).

ture approaches in terms of power. Both the MLMixture and EXppNEX approaches have a maximum difference in power of 0.6% across the four sub-cases. However, at power of 81.2% under scenario 2(a), the EXppNEX approach is marginally better than

the EXNEX model which fails to reach the nominal level at 78.9%. In cases of homogeneity to the current data, the MLMixture and EXppNEX methods give up to a 4.2% increase in power over an EXNEX model, whilst still maintaining error control at or below the nominal value. In fact, the MLMixture model has consistently lower type I error rate than EXppNEX under scenario 2. $\text{EXNEX}_{\text{pool}}$ gives very biased point estimates for the response rates when the historic and current baskets don't align. For instance in scenario 2(a), for basket 1, the current basket is effective to treatment and the historic is not. Under $\text{EXNEX}_{\text{pool}}$, the results are pooled regardless of heterogeneity, resulting in a point estimate of 0.187 when the true response rate is 0.25. A table of point estimates under each method and scenario are provided in Supporting Information C.

Under scenario 5 (presented in the bottom half of Figure 4.4.1), one basket is ineffective against the treatment. This basket has no historic information available. Again, the EXNEX model demonstrates the lowest power as it does not utilise the additional historic information available. For baskets 1-3 with historic information available, the $\text{mEXNEX}_{\text{hist}}$ approach has the highest power in all 4 sub-cases ranging from 90.7-94.4% depending on homogeneity amongst the historic sources with power greater under sub-cases (a) and (d) in which all historic baskets are completely homogeneous. However, this also came with the greatest type I errors in these two sub-cases at around 22%. Inflation in error rates is also an issue in the MLMixture model. The type I error rate across the four sub-cases ranges from 18.2-19.8% under the MLMixture model, a substantial inflation over the nominal 10% level and a substantial increase from the standard EXNEX model whose error rate is 14.5%. In contrast, the EXppNEX approach has reduced error rates compared to the standard EXNEX model ranging from 12.4-14.1% but only shows improvement in power for some baskets, for instance sub-case 5(b) shows the EXppNEX model has an improvement in power in basket 1 at 91.7% compared to 88.5% under the EXNEX model. For the $\text{EXNEX}_{\text{pool}}$ approach, the error rate reduce as the number of effective historic baskets increases, resulting in

a type I error 2.6% lower than the standard EXNEX model under scenario 5(d).

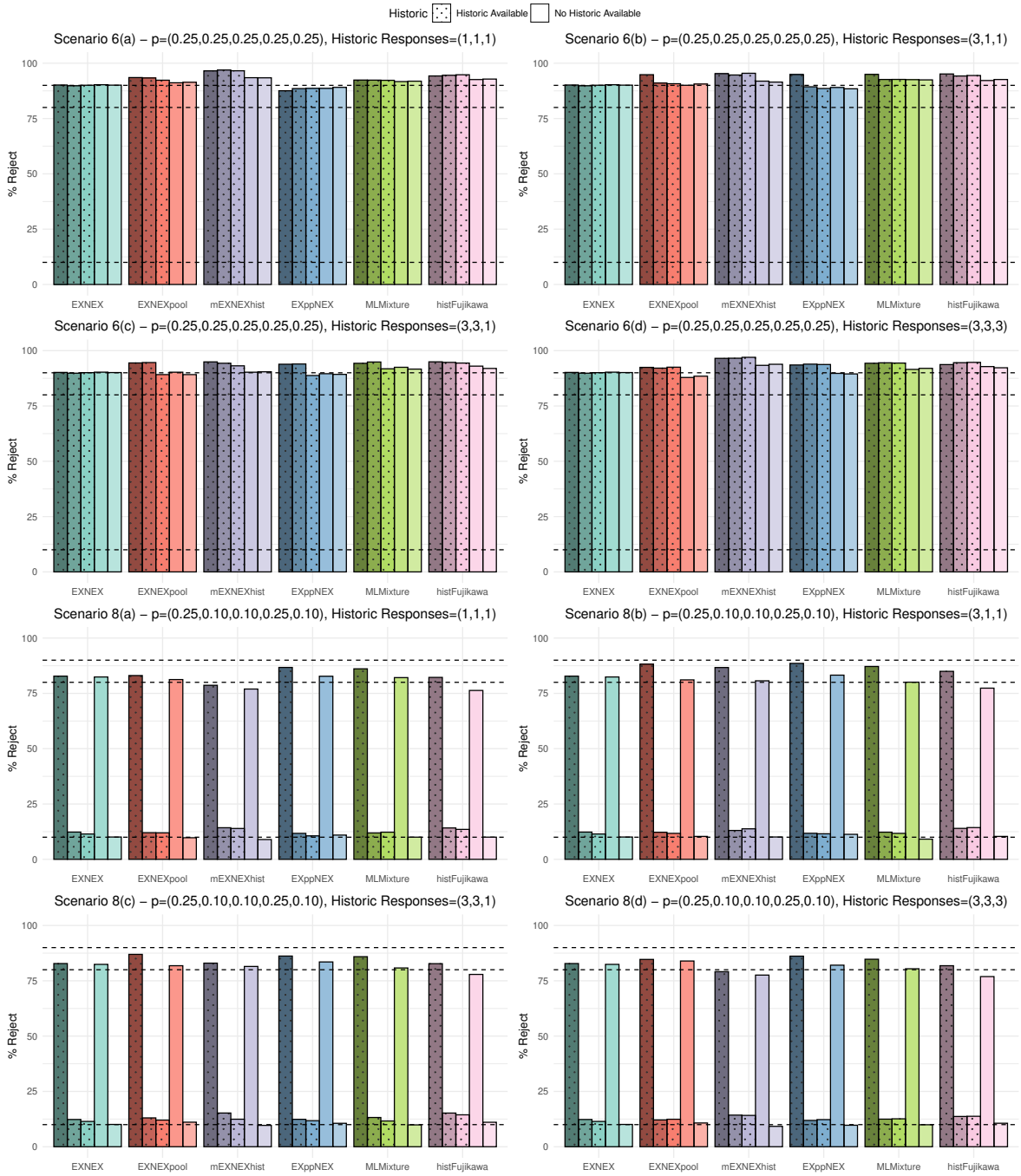


Figure 4.4.2: Type I error rate and power under each of the 6 approaches for historic information borrowing for scenarios 6 and 8 cases (a)-(d).

Figure 4.4.2 presents the results under scenarios 6 and 8. Under scenario 6 all current baskets are homogeneous and effective to treatment. Under sub-case (a) in

which the historic responses are all heterogeneous to the current baskets, it is observed that the $\text{EXNEX}_{\text{pool}}$ gives power of approximately 93.1% for baskets 1-3 with historic information and 91.2% for baskets 4 and 5. However, under sub-case (d) where all historic responses are homogeneous to the current response data, these power values actually decrease to around 92.4% and 88.2% for baskets with and without historic information respectively. This may seem counter intuitive given that the $\text{EXNEX}_{\text{pool}}$ approach borrows fully from the historic information by pooling the results within baskets, so in cases of homogeneity one would expect a further increase in power. However, this is not observed due to the calibration of the efficacy decision criteria. Under sub-case (a), the cut-off value for baskets 1-3 is 0.826 which is far less conservative than the calibrated value obtained under sub-case (d) at 0.966. This is due to the pooling itself, under sub-case (a), the three ineffective historic baskets will always pull the posteriors down in the $\text{EXNEX}_{\text{pool}}$ model, reducing the chance of a type I error, therefore requiring a less stringent cut-off value. This lower Δ_k value also makes it easier to reject the null, thus improving the power. The higher Δ_k value in sub-case (d) results in a reduction of power. This highlights the importance of calibration and how, if done correctly, it can drastically change results of any studies.

Both $\text{mEXNEX}_{\text{hist}}$ and histFujikawa demonstrate the greatest power under scenario 6. For example, under sub-case (d), the $\text{mEXNEX}_{\text{hist}}$ approach has average power of 96.7% for baskets with historic information and 93.7% for those without. This compares to an average power of 90.1% under the EXNEX model. The MLMixture appears to handle heterogeneity between historic sources better than the EXppNEX model in terms of power, giving values consistently above 90% in all sub-cases. The EXppNEX approach has lower power than the EXNEX model under (a) at around 88.5%. This reduction is maintained for all baskets and sub-cases in which the historic and current data are in conflict. However, in cases of homogeneity, power can be substantially increased, for instance under 6(b) EXppNEX has power 94.88% for basket 1.

Table 4.4.4: The average power and maximum type I error rate, computed across the 8 scenarios under all 4 historic data sub-cases. Note that the average is only taken across baskets of the same type i.e. with or without historic baskets and only between baskets with an identical number of responses in the historic basket.

				Average Power			
Sub-Case	**Basket(s)**	EXNEX	$\text{EXNEX}_{\text{pool}}$	$\text{mEXNEX}_{\text{hist}}$	EXppNEX	MLMixture	histFujikawa
(a)	1,2,3	86.65	88.53	88.44	87.20	88.53	87.66
	4,5	86.28	86.28	83.56	85.79	86.81	81.74
(b)	1	85.05	89.64	89.99	89.24	88.82	87.65
	2,3	88.03	89.00	91.00	88.25	89.57	89.96
	4,5	86.28	86.03	86.04	85.90	86.37	82.37
(c)	1,2	86.00	89.07	88.60	88.21	88.91	87.27
	3	88.83	88.08	90.99	88.89	89.87	91.53
	4,5	86.28	85.89	85.52	86.15	86.24	83.61
(d)	1,2,3	86.65	88.11	88.69	88.91	88.59	87.77
	4,5	86.28	86.23	84.17	86.02	86.10	83.04
				Maximum Type I Error Rate			
(a)	1,2,3	12.32	13.66	16.02	12.08	12.26	14.18
	4,5	14.48	17.64	22.08	12.42	18.34	20.46
(b)	1	7.08	11.88	11.62	11.58	11.26	12.90
	2,3	12.32	12.72	13.78	11.76	12.52	14.34
	4,5	14.48	15.84	17.04	13.04	18.58	20.16
(c)	1,2	12.32	13.04	15.24	12.40	13.22	15.22
	3	11.44	12.04	13.76	11.76	12.64	14.42
	4,5	14.48	14.10	16.28	13.70	19.78	19.10
(d)	1,2,3	12.32	12.42	16.34	12.26	12.68	13.76
	4,5	14.48	11.84	22.00	14.10	18.16	19.66

The benefits of utilizing information borrowing from both current and historic sources are highlighted in scenario 6(d) in which all current and historic baskets are homogeneous and effective to the treatment. In this case, all of the proposed methods demonstrate substantial power gain compared to the EXNEX model.

Finally, in scenario 8 one basket without historic information is effective whilst one basket with historic information is also effective, with the rest ineffective to treatment. The results in Figure 4.4.2 show similar findings to those in scenarios 2, 5 and 6 as previously discussed. Across all four sub-cases, of the proposed approaches, the EXppNEX model has highest power for both baskets 2 and 4 whilst maintaining a type I error rate close to that of the EXNEX model. The MLMixture model achieves similar power to the EXppNEX model, however demonstrates greater error inflation with maximum error of 13.2%. For comparison the EXNEX model has maximum error of 12.3%. The $\text{EXNEX}_{\text{pool}}$ model provides reasonable power, however, in cases (a), (b) and (c), this power is reduced compared to the EXNEX model for basket 4. This approach also demonstrates error inflation particularly in the cases in which the historic information is heterogeneous to the current data.

Table 4.4.4 presents the maximum type I error rate across the 8 data scenarios for each approach, split by sub-case and by basket alongside the average power across the data scenarios. It is first observed that type I error inflation is far more substantial in baskets without historic information under both the $\text{mEXNEX}_{\text{hist}}$ and histFujikawa approaches, as well as the MLMixture model. The EXppNEX has far better error control across all five baskets in the trial with a maximum of just 14.1% which occurs in sub-case (d) for baskets 4 and 5. Under this sub-case, the MLMixture model gives maximum error of 18.2%, $\text{mEXNEX}_{\text{hist}}$ 22% and histFujikawa 19.7%. This reduction in error rate under the EXppNEX approach does come alongside a reduction in power compared to the EXNEX model in a handful of cases, however, this reduction does not exceed 0.5%. That being said, power is improved in several cases, with power

improving up to 4.2% over the EXNEX model. The MLMixture model produces similar average power to the EXppNEX approach, improving over the EXNEX model by up to 3.8%, however, has substantially increased error rates in baskets 4 and 5. From these tables, we also observe that the performance of $\text{mEXNEX}_{\text{hist}}$ and histFujikawa fluctuated substantially between baskets and sub-cases, with $\text{mEXNEX}_{\text{hist}}$ giving the highest average power under sub-case (b) but substantially reduced power for baskets 4 and 5 under sub-case (d). The $\text{EXNEX}_{\text{pool}}$ model presented similar average power values to the EXppNEX approach, with a higher power in a number of sub-cases, however, error rates were consistently higher.

To summarise, given these results, the EXppNEX model would be recommended due to its superior error control compared to the other approaches considered in this chapter, including the standard EXNEX model. The EXppNEX model also substantially improves power in cases of homogeneity between historic and current data sources compared to the EXNEX model. However, only a single value of power parameter α was considered above. As stated throughout the literature, there is difficulty surrounding the selection of α, as operating characteristics can be highly dependent on the value (Duan et al., 2006). A sensitivity analysis on the α parameter is presented in Section 4.4.5. Both the $\text{mEXNEX}_{\text{hist}}$ and histFujikawa models are not recommended due to their inconsistent performance, showing substantially decreased power compared to the nominal level under several scenarios. The MLMixture is far more computationally intensive without substantially improving performance in both power and type I error rate compared to the alternative approaches. Although results were only presented here for half of the 32 total scenarios considered, results of the remaining 16 proved similar.

4.4.5 Sensitivity

Although the results in Section 4.4.4 highlighted fairly substantial error inflation under the MLMixture compared to the standard EXNEX model, this error inflation can be

shown to be limited by adjusting the mixture weights $\pi_{\lambda,k}$, $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$. Table 4.4.5 summarises the operating characteristics across the 8 scenarios under several combinations of mixture weights, with $\pi_{\lambda,k}$, $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$ taking values 0.25, 0.5 or 0.75. Several settings of these mixture weights were considered and the maximum type I error rate and average power was taken across all 8 scenarios split by basket and by historic sub-case. Both $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$ are set as equal, thus the mixture weights in the two EXNEX models in the MLMixture model follow the same distribution.

Setting all mixture weights to 0.25 showed a reduction in maximum type I error rate (maximum error rate is 12.5%) in all baskets and historic sub-cases compared to both settings in which $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$ are set to 0.75 (maximum type I error rates of 22.6% and 23.3%). Placing a $\pi_{\text{curr},i} = \pi_{\text{all},i} = 0.75$ weight increases the probabilities of being in both EX components, therefore encouraging borrowing between baskets. In cases of heterogeneity this increased borrowing results in more substantial error inflation over the nominal level. A lower maximum error rate is also observed when using 0.25 for all mixture weights compared to when mixture weights are set to 0.5 (maximum type I error rate of 19.8%) in all bar one setting .

Using equal weights of 0.25 across all mixtures gives a reduction in power compared to the setting where $\pi_{\lambda,k} = 0.25$ and $\pi_{\text{curr},i} = \pi_{\text{all},i} = 0.75$. The maximum difference in power is a reduction of 3.4%, however an increase of up to 2.4% in power is observed in some baskets. Both cases in which $\pi_{\text{curr},i} = \pi_{\text{all},i} = 0.25$ produce similar maximum type I error rate and average power regardless of the choice of $\pi_{\lambda,k}$.

Using equal weights of 0.5 for all mixtures balances the error control observed under the equal mixture weights of 0.25 with the improved power of setting $\pi_{\text{curr},i} = \pi_{\text{all},i} = 0.75$. Results of the full simulation study exploring weights in the MLMixture model are presented in Section C.6 of the Supporting Information C.

Table 4.4.5: A comparison of operating characteristics where weights $\pi_{\lambda,k}$, $\pi_{\text{curr},j}$ and $\pi_{\text{all},j}$ are altered to either 0.25, 0.5 or 0.75. Each setting is labelled as x, y which correspond to a setting of MLMixture weights where x is the value of $\pi_{\lambda,k}$ (set at 0.25, 0.5 or 0.75) and y are the values of $\pi_{\text{curr},j}$ and $\pi_{\text{all},j}$ which are set as equal and to either 0.25, 0.5 or 0.75. The maximum type I error rate (E) and average power (P) are computed across the 8 scenarios under all 4 historic data sub-cases. Note that the maximum/average is only taken across baskets of the same type i.e. with or without historic baskets and only between baskets with an identical number of responses in the historic basket.

Sub-Case	**Basket(s)**	**Weights**									
		0.25,0.25		**0.25,0.75**		**0.75,0.25**		**0.75,0.75**		**0.50,0.50**	
		E	P	E	P	E	P	E	P	E	P
(a) $y_{k^*} = (1,1,1)$	1,2,3	12.12	87.33	14.06	88.34	12.00	86.73	13.48	88.72	12.26	88.53
	4,5	11.86	86.04	22.40	84.43	11.74	86.35	22.86	84.61	18.34	86.81
(b) $y_{k^*} = (3,1,1)$	1	11.46	87.70	12.56	89.77	11.36	87.17	12.19	89.46	11.26	88.12
	2,3	11.84	88.30	13.74	90.41	12.10	87.12	13.56	90.32	12.52	89.57
	4,5	11.72	87.00	22.60	84.69	12.14	85.43	23.34	84.42	18.58	86.37
(c) $y_{k^*} = (3,3,1)$	1,2	11.82	87.05	14.20	88.48	12.44	87.20	15.02	88.24	13.22	88.91
	3	12.46	88.49	13.90	91.86	10.74	88.55	15.56	92.16	13.28	90.69
	4,5	12.16	86.22	23.26	83.78	12.52	86.24	21.60	84.65	19.78	86.24
(d) $y_{k^*} = (3,3,3)$	1,2,3	12.36	87.88	14.02	88.61	12.18	87.49	13.72	88.65	12.58	88.59
	4,5	12.04	85.58	22.26	84.62	12.00	86.06	22.06	83.98	18.16	86.10

Table 4.4.6: A comparison of operating characteristics where power parameter, α, are altered to either 0.25, 0.5 or 1 in the EXppNEX approach. The maximum type I error rate (E) and average power (P) are computed across the 8 scenarios under all 4 historic data sub-cases. Note that the maximum/average is only taken across baskets of the same type i.e. with or without historic baskets and only between baskets with an identical number of responses in the historic basket.

		α					
Sub-Case	**Basket(s)**	**0.25**		**0.5**		**1**	
		E	P	E	P	E	P
(a) $y_{k^*} = (1,1,1)$	1,2,3	11.92	87.49	12.08	87.20	11.98	87.18
	4,5	13.12	86.17	12.42	85.79	12.44	85.88
(b) $y_{k^*} = (3,1,1)$	1	11.84	88.25	11.58	89.24	11.86	88.49
	2,3	12.04	88.48	11.76	88.25	12.14	87.77
	4,5	14.32	86.18	13.04	85.90	13.16	85.66
(c) $y_{k^*} = (3,3,1)$	1,2	12.84	88.40	12.40	88.21	12.26	88.69
	3	12.12	89.12	11.76	88.89	11.76	88.60
	4,5	13.64	86.37	13.70	86.15	12.92	86.05
(d) $y_{k^*} = (3,3,3)$	1,2,3	12.52	88.84	12.26	88.91	12.38	88.68
	4,5	14.32	86.39	14.10	86.02	13.52	86.31

Similarly, alternative choices for the power, α, in the EXppNEX approach were also explored. In the simulation study presented in Section 4.4, α is set to allow for a moderate amount of borrowing. Three alternative values of α were considered: 0.25, 0.5 and 1. A choice of $\alpha = 1$ encourages full borrowing from the historic data in the NEX component, whilst 0.25 discounts the historic data heavily. This simulation study has the same setting as in Section 4.4, with only the power parameter varied. The maximum type I error rate and average power across the 8 scenarios are presented in Table 4.4.6, split by historic sub-case and basket. All power values are consistent across all choices of α, ranging by no more than 1%. Slightly more variation is observed in

the maximum type I error rate, with $\alpha = 0.25$ giving marginally higher error rates in almost all baskets and sub-cases. A choice of $\alpha = 0.5$ results in the smallest error in almost all cases, with an improvement in power compared to $\alpha = 1$ in most cases too. Full results of this simulation study are presented in Section C.5 in Supporting Information C.

4.5 Discussion

In this chapter, we present several approaches for borrowing from both historic and current baskets under one framework. Most proposed approaches built on the exchangeability-nonexchangeability model which has previously been implemented to borrow information between baskets on the current trial. This model was used as a basis due to its popularity in the field of basket trials and due to its flexible structure in terms of allowing both borrowing and an independent analysis in one model. The other proposed method utilised Fujikawa's design and the simple adaptation of incorporating the historic data alongside current data in the posterior parameters proved advantageous due to its reduced computation time. A comparison of computational time of all proposed approaches is presented in the Section C.3 of Supporting Information C, which shows how much more computationally intensive the hierarchical modelling approaches are, while demonstrating that the MLMixture model becomes quickly infeasible to conduct large-scale simulations as the number of current and historic baskets increases.

The conclusion of the simulation study presented in this chapter favoured the use of the EXppNEX approach, however, it is stated throughout the literature that the performance of a power prior is sensitive to the choice of power parameter, α. This was explored in another simulation study with results provided in Section 4.4.5 and in Supporting Information C, where results demonstrated that in this simulation setting, the choice of power parameter had minimal effect on operating characteristics. Alternative

approaches such as the modified power prior (Duan et al., 2006), calibrated power prior (Zheng and Wason, 2022) or the commensurate prior (Hobbs et al., 2011) could also be implemented in place of the power prior in the EXppNEX approach, however, were not considered in this chapter.

Throughout all simulation studies presented in this chapter, the Robust calibration procedure (RCaP) was implemented in order to calibrate efficacy cut-off values. Should calibration have been conducted under the traditional approach of calibrating under the null, more substantial error inflation would have been observed and with that a greater improvement in power would also be present, however, the comparison of approaches remained the same, with differences between their performance slightly more pronounced. The calibration of these efficacy cut-off values is a key component to any simulation study and heavily impact operating characteristics. This is evident in some of the results observed in this chapter, particularly when comparing the EXNEX and $\text{EXNEX}_{\text{pool}}$ approaches as their cut-off values, Δ_k, varied so much in their conservative nature.

This work could be extended further to increase the sensitivity of each of the proposed approaches to the presence of heterogeneity between current baskets but also heterogeneity to the historic data sources. This would hopefully improve the inflation of error rates under some approaches. In particular, the MLMixture as it stands does not control error rates to the nominal level in cases of heterogeneity across all baskets current and historic. Simulation studies found that more weight was placed on the informative NEX prior than was desirable, so altering the form of this prior or adjusting mixture weights could potentially prove beneficial for error control. The simulation study presented in this chapter assumed equal sample sizes for both current and historic studies, as well as, only a single source of historic information for three of the five baskets. Further simulations could be conducted to alter these design parameters to investigate the performance in the case of variability between baskets.

Finally, this chapter focused on a simulation setting motivated by the MyPathway and VE-BASKET trials, with sample sizes of 34 and 13 used in the current and historic baskets respectively. These were based on the planned sample sizes for both studies. A sample size of 34 patients is rather large for a basket trial which typically studies rare diseases. In fact, in the MyPathway study, in the BRAFV600 mutation branch of the trial, not a single basket achieved this planned sample size, with the greatest observed in the NSCLC basket consisting of 14 patients. The use of the larger sample size in the simulation study down-plays the benefits of borrowing from the historic information, as smaller baskets benefit more greatly from this additional source of information. To address this, a further simulation study was conducted with the sample size in the current study reduced from 34 to 20 patients with all other design parameters kept the same. Results demonstrate comparable findings between the performance of methods as presented in this chapter, however, due to the smaller sample size, the nominal power value is rarely reached. In fact, the nominal power of 80% is not achieved using the standard EXNEX model in which historic data is ignored, thus encouraging the use of the proposed historical borrowing techniques.

4.6 Appendix

4.6.1 Models

Presented below are the full model specifications for the simulation study presented in Section 4.4, including all parameter choices and prior specifications. Note that for this simulation study $H_k = 1$ for $k = 1, 2, 3$ and $H_k = 0$ for $k = 4, 5$ and thus only one source of historic information was included, allowing the superscript j to distinguish the historical study, to be dropped.

1. **EXNEX:**

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1, 2, 3, 4, 5, \\
&p_k = \delta_k M_{1k} + (1 - \delta_k) M_{2k}, \\
&\delta_k \sim \text{Bernoulli}(\pi_k), \\
&\theta_{1k} = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)} \\
&\mu \sim \text{N}(\text{logit}(0.1), 10^2), \\
&\sigma \sim \text{Half-Normal}(0, 1), \\
&\theta_{2k} = \text{logit}(M_{2k}) \sim \text{N}(-1.386, 6.25^2). \quad \text{(NEX)}
\end{aligned}
$$

   with $\pi_k = 0.5$ for $k = 1, 2, 3, 4, 5$.

2. **EXNEX$_{\text{pool}}$:**

$$
\begin{aligned}
&Y_k = y_k + y_{k^*} \sim \text{Binomial}(n_k + n_{k^*}, p_k), \qquad k = 1, 2, 3, 4, 5, \\
&p_k = \delta_k M_{1k} + (1 - \delta_k) M_{2k}, \\
&\delta_k \sim \text{Bernoulli}(\pi_k), \\
&\theta_{1k} = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)} \\
&\mu \sim \text{N}(\text{logit}(0.1), 10^2), \\
&\sigma \sim \text{Half-Normal}(0, 1), \\
&\theta_{2k} = \text{logit}(M_{2k}) \sim \text{N}(-1.386, 6.25^2). \quad \text{(NEX)}
\end{aligned}
$$

   where $y_{k^*} = n_{k^*} = 0$ for $k = 4, 5$, i.e. in baskets without historic information available. Mixture weights are $\pi_k = 0.5$ for $k = 1, 2, 3, 4, 5$.

3. **mEXNEX$_{\text{hist}}$:** same as the EXNEX model above but with

$$
\pi_k = \sum_{i^*=1, i^* \neq k^*}^{3} \frac{1 - h_{i^*,k^*}}{3 - 1} \quad \text{for } k = 1, 2, 3, \qquad \text{then } \pi_k = 0.8 \sum_{i=1}^{3} \frac{\pi_i}{3} \quad \text{for } k = 4, 5,
$$

where h_{i^*,k^*} is the Hellinger distance between the posteriors of two historic baskets i^* and k^*.

4. **Fujikawa$_{\text{hist}}$:** Each basket has the following closed form posterior

$$\pi(p_k|D, D_h) = \text{Beta}\Bigg(1 + \sum_{i=1}^{5}(\mathbb{I}\{\omega_{k,i}^{\epsilon} > \tau\}\omega_{k,i}^{\epsilon} y_i + 0.8\mathbb{I}\{\omega_{k,i^*}^{*\epsilon} > \tau\}\omega_{k,i^*}^{*\epsilon} y_{i^*}),$$
$$1 + \sum_{i=1}^{5}(\mathbb{I}\{\omega_{k,i}^{\epsilon} > \tau\}\omega_{k,i}^{\epsilon}(n_i - y_i) + 0.8\mathbb{I}\{\omega_{k,i^*}^{*\epsilon} > \tau\}\omega_{k,i^*}^{*\epsilon}(n_{i^*} - y_{i^*}))\Bigg)$$

where $\omega_{k,i} = 1 - \text{JSD}(\pi(p_k|Y_k = y_k), \pi(p_i|Y_i = y_i))$ and

$$\omega_{k,i^{*(j)}} = \begin{cases} 1 - JSD\left(\pi(p_k|Y_k = y_k), \pi(p_{i^*}^{(j)}|Y_{i^*}^{(j)} = y_{i^*}^{(j)})\right) & \text{If historic information,} \\ & \text{is available for basket } i \\ 0 & \text{Otherwise.} \end{cases}$$

with $\epsilon = 2$ and $\tau = 0.2$.

5. **EXppNEX:**

$$Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1, 2, 3, 4, 5,$$
$$p_k = \delta_k M_{1k} + (1 - \delta_k) M_{2k},$$
$$\delta_k \sim \text{Bernoulli}(\pi_k),$$
$$\mathbb{I}_k = 1 \text{ if } y_{k^*} \text{ exists for basket } k,$$
$$\theta_{1k} = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2), \quad \text{(EX)}$$
$$\mu \sim \text{N}(\text{logit}(0.1), 10^2),$$
$$\sigma \sim \text{Half-Normal}(0, 1),$$
$$M_{2k} = \mathbb{I}_k P_{1k} + (1 - \mathbb{I}_k) P_{0k},$$
$$P_{1k} \sim \text{Beta}(1 + \alpha y_{k^*}, 1 + \alpha(n_{k^*} - y_{k^*})),$$
$$\theta_{2k} = \text{logit}(P_{0k}) \sim \text{N}(-1.386, 6.25^2),$$

with $\pi_k = 0.5$ for $k = 1, 2, 3, 4, 5$ and $\alpha = 0.5$ for all baskets.

6. **MLMixture:**

$$Y_i \sim \text{Binomial}(n_i, p_i) \qquad i = 1, 2, 3, 4, 5, 1^*, 2^*, 3^*,$$

$$\psi_i = \begin{cases} 1 & \text{if basket } i \text{ is a historic basket,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\gamma_{\text{all},i} = \epsilon_{\text{all},i}\text{EX}_{\text{all},i} + (1 - \epsilon_{\text{all},i})\text{NEX}_{\text{all},i},$$

$$\epsilon_{\text{all},i} \sim \text{Bernoulli}(\pi_{\text{all},i}),$$

$$\text{EX}_{\text{all},i} \sim \text{N}(\mu_{\text{all}}, \sigma^2_{\text{all}}),$$

$$\mu_{\text{all}} \sim \text{N}(m_{\mu_{\text{all}}}, \nu_{\mu_{\text{all}}}),$$

$$\sigma_{\text{all}} \sim g(\cdot)$$

$$N_{\text{all},i} \sim \text{Beta}\left(a_i + (1 - \psi_i)\sum_{t=1}^{H_i} y_{i^*(t)}, b_i + (1 - \psi_i)\sum_{t=1}^{H_i}(n_{i^*(t)} - y_{i^*(t)})\right),$$

$$\text{NEX}_{\text{all},i} = \text{logit}(N_{\text{all},i}),$$

$$\text{EXNEX}_{\text{all},i} = \exp(\gamma_{\text{all},i})/(1 + \exp(\gamma_{\text{all},i})),$$

$$\gamma_{\text{curr},i} = \epsilon_{\text{curr},i}\text{EX}_{\text{curr},i} + (1 - \epsilon_{\text{curr},i})\text{NEX}_{\text{curr},i},$$

$$\epsilon_{\text{curr},i} \sim \text{Bernoulli}((1 - \psi_i)\pi_{\text{curr},i}),$$

$$\text{EX}_{\text{curr},i} \sim \text{N}(\mu_{\text{curr}}, \sigma^2_{\text{curr}}),$$

$$\mu_{\text{curr}} \sim \text{N}(m_{\mu_{\text{curr}}}, \nu_{\mu_{\text{curr}}}),$$

$$\sigma_{\text{curr}} \sim f(\cdot)$$

$$N_{\text{curr},i} \sim \text{Beta}(a_i, b_i),$$

$$\text{NEX}_{\text{curr},i} = \text{logit}(N_{\text{curr},i}),$$

$$\text{EXNEX}_{\text{curr},i} = \exp(\gamma_{\text{curr},i})/(1 + \exp(\gamma_{\text{curr},i})),$$

$$p_k = \lambda_k \text{EXNEX}_{\text{all},k} + (1 - \lambda_k)\text{EXNEX}_{\text{curr},k}, \quad k = 1, 2, 3, 4, 5, \tag{4.6.1}$$

$$\lambda_k \sim \text{Bernoulli}(\pi_{\lambda,k}). \tag{4.6.2}$$

All mixture weights are set equal such that $\pi_{\lambda,k} = 0.5$ for all $k = 1, 2, 3, 4, 5$ and $\pi_{\text{curr},i} = \pi_{\text{all},i} = 0.5$ for all $i = 1, 2, 3, 4, 5, 1^*, 2^*, 3^*$.

4.6.2 Calibrated Δ_k Values under the RCaP

Table 4.6.1: Calibrated Δ_k values obtained using the RCaP procedure across the 8 scenarios presented in Table 4.4.1. This is conducted under each of the four historic data settings separately.

Method	$\mathbf{p_1}$	$\mathbf{p_2}$	$\mathbf{p_3}$	$\mathbf{p_4}$	$\mathbf{p_5}$
$\mathbf{y_{k^*}}$	**1**	**1**	**1**		
EXNEX	0.900	0.900	0.900	0.900	0.900
$\text{EXNEX}_{\text{pool}}$	0.826	0.826	0.826	0.902	0.902
$\text{mEXNEX}_{\text{hist}}$	0.914	0.914	0.914	0.935	0.935
histFujikawa	0.967	0.967	0.967	0.985	0.985
EXppNEX	0.897	0.897	0.897	0.889	0.889
MLMixture	0.863	0.863	0.863	0.899	0.899
$\mathbf{y_{k^*}}$	**3**	**1**	**1**		
EXNEX	0.900	0.900	0.900	0.900	0.900
$\text{EXNEX}_{\text{pool}}$	0.944	0.838	0.838	0.905	0.905
$\text{mEXNEX}_{\text{hist}}$	0.857	0.894	0.894	0.909	0.909
histFujikawa	0.972	0.984	0.984	0.993	0.993
EXppNEX	0.909	0.895	0.896	0.890	0.890
MLMixture	0.895	0.875	0.875	0.911	0.911
$\mathbf{y_{k^*}}$	**3**	**3**	**1**		
EXNEX	0.900	0.900	0.900	0.900	0.900
$\text{EXNEX}_{\text{pool}}$	0.955	0.955	0.851	0.912	0.912
$\text{mEXNEX}_{\text{hist}}$	0.891	0.891	0.891	0.912	0.912
histFujikawa	0.991	0.991	0.992	0.997	0.997
EXppNEX	0.917	0.917	0.889	0.884	0.884
MLMixture	0.917	0.917	0.885	0.921	0.921
$\mathbf{y_{k^*}}$	**3**	**3**	**3**		
EXNEX	0.900	0.900	0.900	0.900	0.900
$\text{EXNEX}_{\text{pool}}$	0.966	0.966	0.966	0.918	0.918
$\text{mEXNEX}_{\text{hist}}$	0.912	0.912	0.912	0.933	0.933
histFujikawa	0.996	0.996	0.996	0.999	0.999
EXppNEX	0.916	0.916	0.916	0.879	0.879
MLMixture	0.930	0.930	0.930	0.930	0.930

4.6.3 Simulation Results

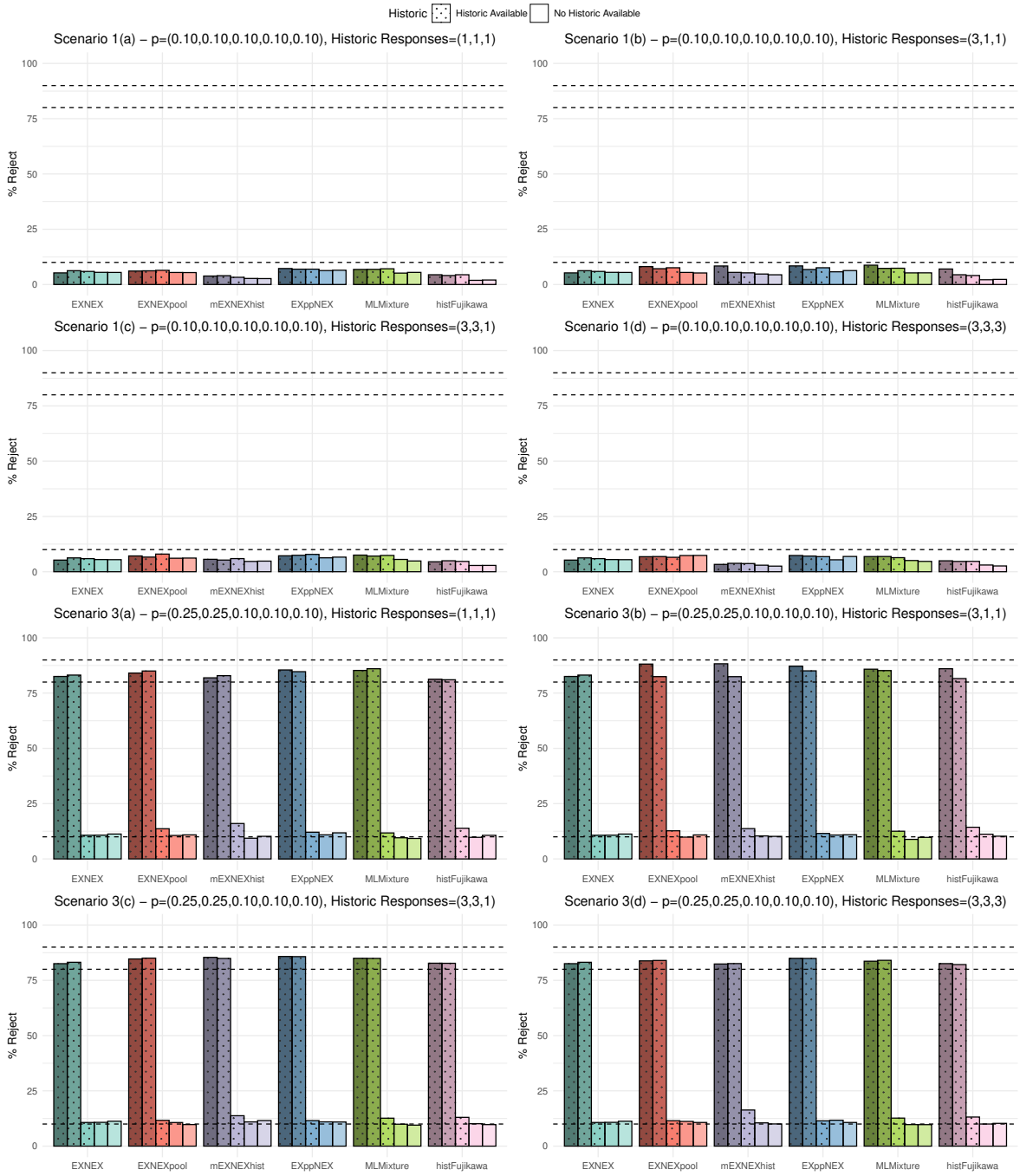


Figure 4.6.1: Type I error rate and power under each of the 6 approaches for historic information borrowing for scenarios 1 and 3 cases (a)-(d).

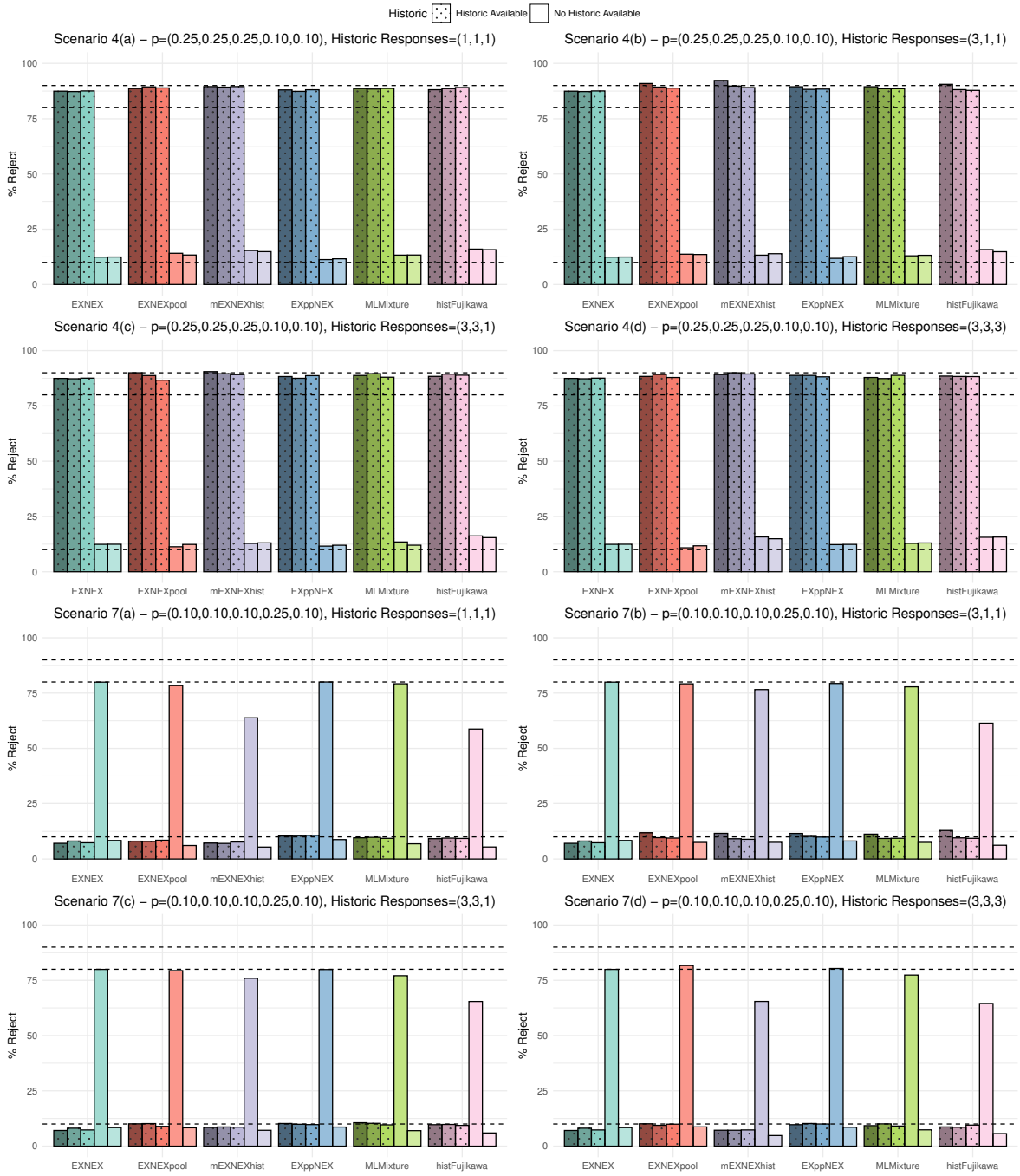


Figure 4.6.2: Type I error rate and power under each of the 6 approaches for historic information borrowing for scenarios 4 and 7 cases (a)-(d).

Chapter 5

Conclusions & Further Research

5.1 Conclusion

This thesis has explored the use of Bayesian methodology in the design and analysis of basket clinical trials in order to aide inference on the efficacy of treatments. Such trials play a key role in the development of treatments for rare conditions, particularly in the oncology setting where sample sizes tend to be limited. The methods proposed in this thesis addresses the issue of a lack of statistical power and precision of estimates in cases where the sample sizes are small. This is achieved through the use of Bayesian information borrowing models, however, as shown throughout this work, such methods come at the risk of an inflation in the basket-wise type I error rate. This thesis explores current literature on the use of Bayesian information borrowing methods in the basket trial setting, proposing novel methodology that improves power and precision of estimates over a stratified analysis of each basket, whilst focusing on type I error control.

In Chapter 2, a detailed comparison is made between various approaches for Bayesian information borrowing in the basket trial setting. This is conducted through extensive simulation studies motivated by the VE-BASKET trial. This study was carried

out under both an equal sample size setting across sub-groups but also considered a more realistic case of unequal sample sizes. These simulation results demonstrated that a Bayesian hierarchical model, an exchangeability-nonexchangeability model and a Bayesian model averaging approach substantially inflate the type I error rate in cases of heterogeneity between baskets responses. The calibrated Bayesian hierarchical model lacks power in such a scenario. This chapter also highlights the inadequacy of the calibration procedure of the CBHM when applied to unequal sample sizes, thus a generalisation of this calibration was made to handle such a setting. To combat the observed type I error inflation, we proposed an adaptation to the EXNEX model, mEXNEX$_c$, which utilises a data-driven approach to set the prior probabilities of exchangeability in the EXNEX model, making it more robust to the presence of heterogeneity. In fact, simulation results demonstrated that, given a suitable choice of tuning parameter, the proposed model possesses better control of the type I error rate, alongside improved power.

Chapter 3 considers two aspects of the design and analysis of basket trials, the first being the addition of baskets to an ongoing trial. In cases of rare diseases, any basket added at a later time will suffer from even smaller sample sizes, thus the role of information borrowing becomes more prominent. Chapter 3 utilises the EXNEX model to propose several approaches for the analysis of new baskets added to the trial. Simulation studies found that no one of the outlined approaches for adding outperformed its competitors across all scenarios. However, as expected, in cases of homogeneity between the new and existing baskets, using stronger information borrowing yields better results, but in cases of heterogeneity to the new basket, analysing it as independent limits error inflation and loss in power. Further simulation studies demonstrated that all approaches are fairly robust to the timing of addition of new baskets, therefore determining that the size of the new basket has no detrimental effect on power and error of baskets that commenced enrolment at the start of the trial.

The second aspect of Chapter 3 is the development of a novel approach for the calibration of efficacy decision criteria. In a basket trial, posterior probabilities are compared to a cut-off value and an efficacy conclusion made. Typically this cut-off value is calibrated under a global null data scenario to control error rates to a nominal level. However, when applied to a non-null scenario, this traditional approach is shown through simulation studies to be inadequate in controlling error inflation when information borrowing methods are used. Thus, an alternative approach is proposed: the robust calibration procedure; which controls for error rates on average across several scenarios. This is a flexible approach, allowing clinicians to specify potential outcomes of the trial in which one would like to control the error rate across, with the ability to weight scenarios that have a higher likelihood of occurring. Compared to the traditional approach, this novel proposal demonstrates superior error control with only a small loss in power relative to the targeted value in a handful of cases.

In Chapter 4, the use of historical or external data is explored to further improve power and precision of estimates. Various approaches are explored for the combination of two forms of information borrowing: borrowing between current baskets and borrowing from historic data under a single analysis framework. Most approaches build on the EXNEX model, with one updating the mixture weights based on the homogeneity between historic baskets, with another incorporating a power prior in the nonexchangeability component to directly borrow from the historic data. A more computationally expensive approach is the proposed multi-level mixture model which consists of a mixture of two EXNEX models, with one containing the historic data. The rationale behind all proposed approaches was to create a data-driven approach which adapts borrowing based not only homogeneity between current baskets but also the homogeneity to historic data. Although all approaches considered demonstrated an improvement in power compared to an EXNEX model which ignores any external data sources, results demonstrate that the addition of the power prior into the nonexchangeability component of

the EXNEX model provided a superior balance of type I error control and improvement in power.

Overall, this thesis has proposed various methods to improve inference within basket trials. The focus was on improving power and precision of estimates whilst pursuing error control. Results of thorough simulation studies have been presented throughout this thesis to demonstrate operating characteristics such as error and power of all discussed methods under a variety of settings and scenarios. All simulation studies have been based on real-life basket trials, most notably the VE-BASKET study, motivating the need for such novel methodology. As stated by Kopp-Schneider et al. (2020), when implementing information borrowing in the case of heterogeneity between data sources, there will be a trade-off of inflation in error and bias along with any potential gain in power. The proposed methods aim to minimise this trade off, using adaptive data-driven techniques for borrowing, as well as, the proposed approach for calibrating efficacy decision criteria. The overall contributions of this thesis is the improvement of inference in basket trials for rare condition, providing patients beneficial treatments in a shorter time frame, as well as, limiting exposure to potentially ineffective treatments.

5.2 Further Work

There are multiple potential areas for further research in the topic of Bayesian methods in basket trials, some of which focuses on the limitations of the models already proposed in this thesis, while others could contribute to the wider field of basket trials.

Chapter 4 explored approaches for utilising historic information in an information borrowing model to further improve power and precision of estimates. One approach outlined was the multi-level mixture model, consisting of a mixture of two EXNEX models, one containing current and historic information and one containing just current baskets. The first component of this model contained an informative prior based on

the historic data. However, through simulation studies it was found that more weight was placed on this component than desirable in the case of heterogeneity between baskets' observed responses. This contributed to some of the observed error inflation. In order to limit the contribution of the historic data within this component in cases of heterogeneity, a power prior or an alternative adaptive approach (such as the calibrated power prior, a robust mixture prior of a meta-analytic prior) could be implemented in the hope to better control error rates.

Furthermore, mixture weights for all methods described in Chapters 3 and 4 could be explored in more depth. As discussed in Chapter 1, the prior probability of exchangeability in the EXNEX model is updated to some degree based on the homogeneity of the data but is not sensitive enough to the heterogeneity of responses. If mixture weights are set at 0.5, as done throughout Chapter 4, it is anticipated that the probability of borrowing from a heterogeneous basket will be too high, which in turn will inflate the type I error rate. Therefore, for all methods in Chapter 4, future work could be conducted to define mixture weights using a data-driven approach to increase them in cases of homogeneity to encourage borrowing and decrease them in cases of heterogeneity, with the goal to further improve error control. This could similarly be explored in the approaches for the addition of a basket as discussed in Chapter 3.

Throughout this thesis adaptive design features such as interim analyses with futility/efficacy stopping have not been considered or incorporated into the simulation studies or methodology. It has been considered in the work by Berry et al. (2013), Chu and Yuan (2018), Jin et al. (2020), and Psioda et al. (2021) and is a key aspect of many modern clinical trials. Such features are desirable as they allow a trial to end earlier should treatments show efficacy, allowing the treatment to be moved to the next stage of drug development at a much quicker rate whilst reducing the number of patients required. Additionally, futility analysis allows treatment arms to be dropped if not showing sufficient efficacy, reducing the number of patients exposed to a potentially

harmful or non-effective treatment. Futility/efficacy stopping could be implemented throughout the simulation settings explored in this thesis and the effect of their inclusion explored. Due to the multiple looks at the data, it is likely that an increase in error rates would occur, however, this could be limited using multiple comparison correction techniques. When incorporating interim analyses, whether to use information borrowing to determine efficacy/futility could be explored as one could debate either for or against its implementation. Another issue to consider is the timing of interim analyses when sample sizes are limited. For example, the VE-BASKET trial consisted of a basket of just 7 patients, therefore, would an interim and decision making be suitable in this case based on just a handful of patients? Similarly, the case of unequal sample sizes and recruitment rates needs to be considered when setting the timing of an interim analysis.

In addition, there is a potential to use data observed at the interim to update the degree of borrowing in the final analysis based on the observed level of heterogeneity. This could be incorporated into prior distributions or the probability of exchangeability in the EXNEX model. Similarly, the degree of heterogeneity between baskets could be used to adjust sample sizes based on interim data. In cases of homogeneity between baskets' response data, it is likely that when using information borrowing, a smaller sample size would be required to achieve the same statistical power as baskets with a larger sample size but with heterogeneity amongst baskets. Similarly, the required sample size could be increased in cases of heterogeneity in order to meet the required power given that information borrowing methods will not increase power in this case. Sample size re-estimation can have a significant impact on the cost and duration of studies (Mano et al., 2023). However, the use in basket trials for rare disease types may be challenging due to the already small sample sizes. Zheng et al. (2023) proposed an approach for determining sample sizes in basket trials where information borrowing is utilised but the inclusion of an interim to adjust these sample sizes was not considered.

Another area for further consideration is a patients basket assignment. Multiple primary cancers occurs when an individual has more than one cancer in the same or different organ (not including cases of metastasis) (Okeke et al., 2023). As discussed by Vogt et al. (2017), the fact that a patient can present with multiple cancers simultaneously is not a new concept but one that is rarely discussed. The challenge in this setting is identifying a treatment strategy that targets both cancer types. To our knowledge, cases of multiple primary cancers within patients have not been discussed in the basket trial setting. This poses a problem as a patients basket assignment is no longer straightforward and a choice must be made to sort the patient into a sub-group. As all patients receive the same treatment, it is unlikely that bias from the choice of basket assignment would be an issue, however, confounding bias may occur. The presence of a secondary cancer may act as a confounding variable when assessing the efficacy of a treatment on the primary cancer, thus treatment responses may differ from patients with a single cancer type. Similarly, the presence of multiple disease types within a patient may create a stronger correlation in responses between baskets, potentially increasing the probability of exchangeability. However, this correlation may not occur between all patients, so one could argue the degree of exchangeability between baskets would decrease. This brings into question how information borrowing would occur in such a scenario and how to determine the degree of borrowing that is implemented. Note, although the above discussed the presence of multiple cancers, the concept can be applied outside of the oncology setting with patients suffering from comorbidities.

This thesis focused solely on binary endpoints, where patients either respond positively to the treatment or not. Interest lay in estimating the response rate in each basket. A natural extension of this is to consider non-binary endpoints, for example a continuous endpoint following a normal distribution, or in the case of randomised basket trials, the treatment effect, i.e. the difference between the experimental and control arms. As stated by Shahapur et al. (2022), in oncology trials, the primary

endpoint of interest is often the overall survival rate (time from enrolment to death) with secondary endpoints including quality of life and tumour-specific endpoints. Shahapur et al. (2022) provide a summary of endpoints used in the oncology setting. The drawback of using overall survival as the primary endpoint is the sheer length of time it takes to observe survival particularly in slow progressing conditions, thus surrogate endpoints are often used in their place such as progression-free survival, which assesses time from enrolment to the first evidence of disease progression. This can be observed quicker than overall survival.

In a single trial several endpoints may be of interest in order to determine a treatments efficacy. The use of multiple endpoints within basket trials, and in particular Bayesian information borrowing frameworks requires further research. Composite endpoints combine several endpoints of interest into a single quantity (Baracaldo-Santamaría et al., 2023) and are typically implemented to reduce follow-up periods and cost. As they are a single quantity, information borrowing between such endpoints is fairly straightforward to implement, however interpretation may be challenging. In particular, some endpoints may have higher importance than others and when making inference it may be of interest to determine the contribution of these endpoints to the overall conclusion. In an information borrowing structure, the exchangeability between different endpoints may vary. Thus, it could be beneficial to develop methodology which can incorporate several endpoints into a single model with varying degrees of borrowing implemented for each based on observed heterogeneity. It may also be feasible to weight endpoints based on importance to the overall efficacy conclusion.

Finally, in order for the work in this thesis to be implemented in real-life trials, it is vital that clinicians have access to the resources to implement the methodology outlined. Several R packages have been produced for basket trial analyses, one of which is the 'bhmbasket' package (Wojciekowski, 2022) which evaluates binary response data from basket trials under several information borrowing models including the BHM and

EXNEX models. Exploratory simulation studies can be conducted within the package to explore operating characteristics. The development of a similar package could be done to ease the implementation of the approaches outlined in this thesis, in particular, in the work outlined in Chapter 4 where historical information is incorporated through more complex methodology.

The work in this thesis covers a single area of the vast field of adaptive clinical trials and personalised medicine. There has been, and will continue to be evolution in clinical trial designs to keep up with the developing needs for efficient studies for a variety of health conditions. This was particularly evident in the COVID-19 pandemic, which will continue to shape and influence the implementation of master protocols in clinical trials. With the rapid increase in popularity in basket trials, further interesting statistical challenges are likely to arise, however, we hope that this thesis will aide in promoting the use of Bayesian methodology in such studies to improve trial outcomes.

Appendix A

Supporting Information: A Comparison of Bayesian Information Borrowing Methods in Basket Trials and a Proposal of Modified EXNEX Method

A.1 Simulation Results for Section 2.3

Table A.1.1: Calibrated Δ_α values for the simulation study in Chapter 2 based on a planned sample size of 13 per basket. These cut-offs are also applied to the realised sample size scenario without re-calibration.

	Δ_α
Independent	0.905
BHM	0.831
CBHM	0.907
BMA	0.871
EXNEX	0.868
$\text{mEXNEX}_{1/13}$	0.880
mEXNEX_0	0.920

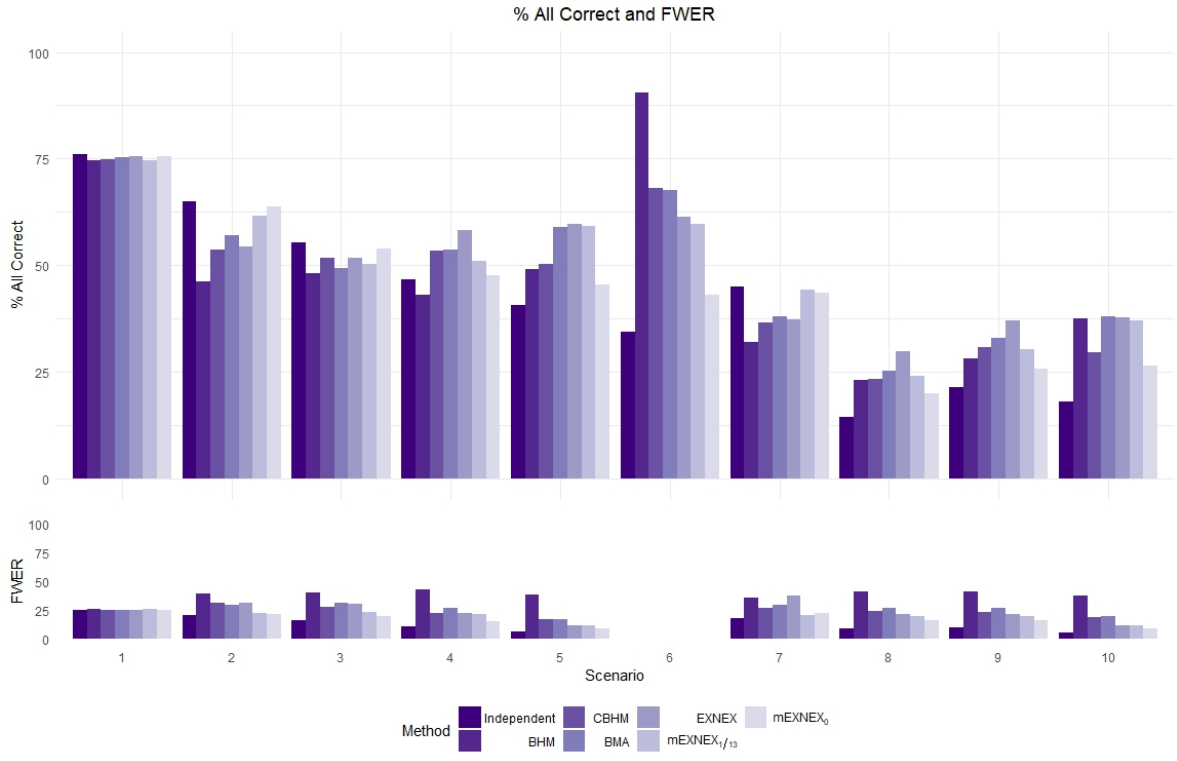


Figure A.1.1: Simulation results for Chapter 2: The family-wise error rate (FWER) and percentage of simulated data sets within which correct inference is made across all baskets (% All Correct) for each method under each data scenario based on a planned sample size of 13 patients per basket.

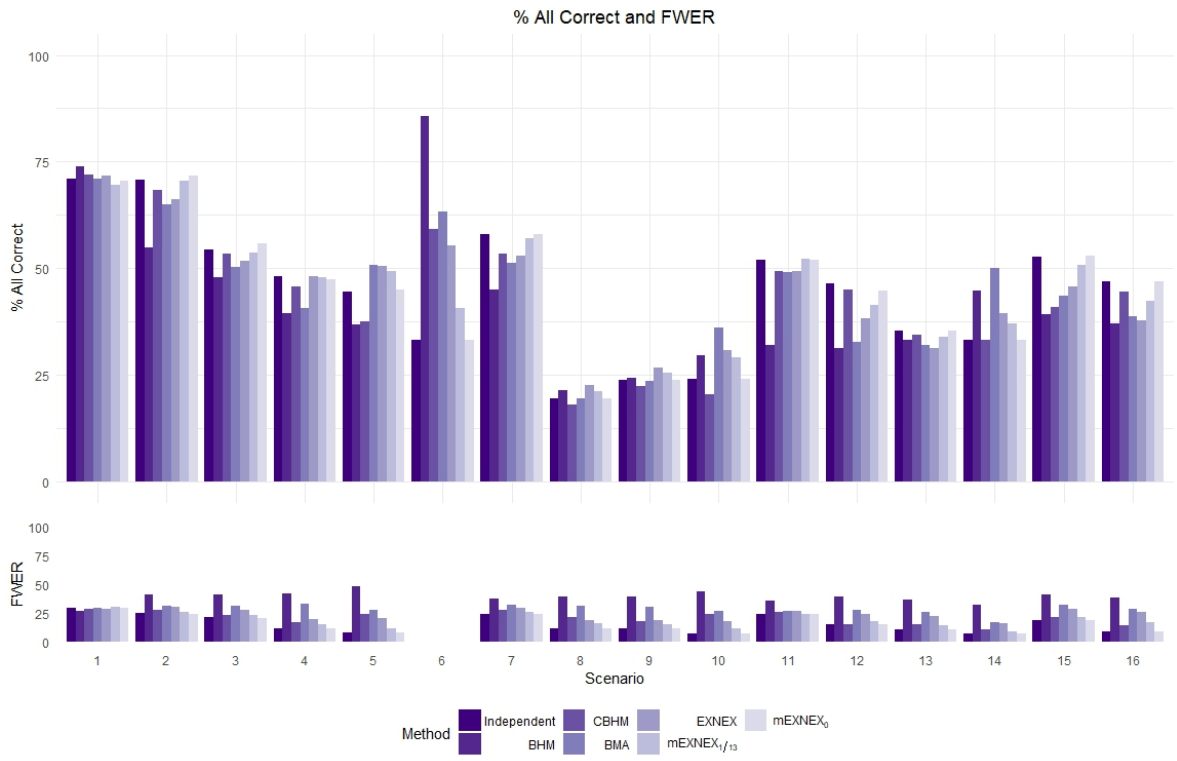


Figure A.1.2: Simulation results for Chapter 2: The family-wise error rate (FWER) and percentage of simulated datasets in which the correct inference is made across all baskets (% All Correct) for each method under each data scenario based on realised sample sizes of 20, 10, 8, 18 and 7 across the 5 baskets.

Table A.1.2: Simulation results for Chapter 2: Operating characteristics for a simulation based on the planned sample size of 13 per basket for scenarios 1-6

.

	% Reject					% All Correct	FWER
Sample Size	13	13	13	13	13		
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	9.72	9.67	10.04	10.22	10.44	58.73	0.413
BHM	9.42	9.52	9.52	9.29	9.51	72.23	0.278
CBHM	9.77	9.93	9.60	9.65	9.90	76.63	0.234
BMA	10.07	10.04	9.78	10.21	9.73	68.05	0.320
EXNEX	10.35	9.95	10.17	9.97	10.35	62.69	0.373
$\text{mEXNEX}_{1/13}$	9.78	10.05	10.40	10.00	10.09	61.18	0.388
mEXNEX_{0}	9.57	10.06	9.18	9.84	9.61	60.04	0.400
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	87.34	10.41	10.61	10.08	10.08	56.79	0.351
BHM	85.51	16.53	16.82	17.16	17.12	45.72	0.419
CBHM	81.13	9.68	9.86	9.82	9.57	56.15	0.275
BMA	86.40	13.16	12.98	12.92	13.59	53.25	0.356
EXNEX	86.89	11.36	12.04	11.99	11.71	51.47	0.387
$\text{mEXNEX}_{1/13}$	87.81	11.37	11.83	11.27	11.67	54.37	0.369
mEXNEX_{0}	87.97	10.35	10.39	10.17	10.63	56.29	0.352
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**		
Independent	88.36	88.67	10.24	9.80	9.97	57.35	0.271
BHM	91.62	91.56	21.70	21.59	22.32	45.96	0.428
CBHM	84.63	84.38	10.44	10.79	10.66	52.99	0.246
BMA	89.93	89.67	17.96	18.33	18.66	49.36	0.358
EXNEX	89.92	90.00	12.55	12.97	12.79	55.04	0.321
$\text{mEXNEX}_{1/13}$	89.53	89.20	12.56	12.41	12.59	54.24	0.316
mEXNEX_{0}	88.39	88.69	10.44	10.59	10.92	55.60	0.282
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**		
Independent	88.37	87.95	88.29	10.37	9.62	56.03	0.189
BHM	94.19	94.03	93.90	29.67	30.44	41.43	0.458
CBHM	85.94	86.48	86.22	12.29	12.06	49.82	0.200
BMA	92.38	92.72	93.09	23.80	23.24	44.18	0.390
EXNEX	91.13	91.08	90.96	13.12	13.13	57.75	0.230
$\text{mEXNEX}_{1/13}$	90.98	90.78	90.68	13.84	14.05	56.37	0.243
mEXNEX_{0}	89.48	89.20	88.89	10.61	10.86	56.53	0.204
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**		
Independent	88.30	87.81	87.46	88.51	10.10	54.14	0.101
BHM	96.55	96.13	96.44	96.04	42.11	47.17	0.421
CBHM	87.99	87.94	87.89	88.19	16.65	47.18	0.167
BMA	94.69	94.44	94.92	94.62	24.01	60.37	0.240
EXNEX	91.28	91.14	91.70	90.91	16.12	56.87	0.161
$\text{mEXNEX}_{1/13}$	91.43	91.72	91.63	91.52	14.86	58.60	0.149
mEXNEX_{0}	89.54	89.44	89.40	89.19	11.20	56.54	0.112
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	88.28	87.66	88.16	88.09	87.95	52.77	
BHM	97.94	98.28	98.13	98.23	97.87	91.53	
CBHM	91.28	91.48	91.06	91.16	91.48	66.72	
BMA	95.24	95.49	95.62	95.30	95.84	79.88	
EXNEX	92.42	92.61	91.98	91.96	91.98	68.39	
$\text{mEXNEX}_{1/13}$	91.64	92.13	92.06	91.60	91.94	66.38	
mEXNEX_{0}	89.78	89.57	89.65	89.92	89.92	58.38	

Table A.1.3: Simulation results for Chapter 2: Operating characteristics for a simulation based on the planned sample size of 13 per basket for scenarios 7-10

.

	% Reject					% All Correct	FWER
Sample Size	13	13	13	13	13		
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	66.04	9.11	9.06	9.03	9.21	45.47	0.317
BHM	63.60	15.13	15.31	15.17	15.16	32.43	0.375
CBHM	55.78	9.13	9.22	8.96	9.26	38.29	0.234
BMA	65.20	13.14	13.37	12.89	13.26	37.72	0.353
EXNEX	67.77	11.27	11.46	11.25	11.31	40.06	0.371
$\text{mEXNEX}_{1/13}$	68.34	11.23	11.02	10.90	11.17	42.22	0.353
mEXNEX_{0}	68.34	10.12	9.84	9.85	10.16	44.90	0.353
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	66.80	64.94	65.95	9.15	9.32	23.70	0.177
BHM	80.04	79.02	80.02	28.93	28.91	24.22	0.426
CBHM	66.33	65.28	65.86	16.20	16.22	20.62	0.225
BMA	77.39	76.45	77.23	23.73	23.78	19.52	0.391
EXNEX	73.23	71.79	73.12	13.39	13.71	28.66	0.231
$\text{mEXNEX}_{1/13}$	72.98	71.60	72.82	14.40	14.43	27.53	0.245
mEXNEX_{0}	69.35	67.97	69.41	10.73	10.84	26.76	0.203
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	86.79	65.64	66.13	9.18	8.98	31.18	0.173
BHM	93.78	79.49	80.73	29.23	29.07	28.67	0.435
CBHM	86.56	65.29	66.40	14.29	14.42	27.79	0.213
BMA	92.90	76.68	77.27	24.01	24.15	25.25	0.395
EXNEX	90.94	71.88	73.18	13.35	13.39	36.03	0.230
$\text{mEXNEX}_{1/13}$	90.71	71.74	73.02	13.83	13.85	35.40	0.238
mEXNEX_{0}	88.57	68.31	70.05	10.96	10.86	34.07	0.205
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**		
Independent	87.23	86.71	66.48	66.23	9.11	30.24	0.091
BHM	95.87	95.57	86.07	86.09	40.76	35.86	0.408
CBHM	88.69	87.89	69.31	69.35	18.48	26.39	0.185
BMA	94.97	94.81	83.30	83.24	27.25	46.40	0.273
EXNEX	91.56	90.86	74.23	74.02	17.25	34.31	0.173
$\text{mEXNEX}_{1/13}$	91.59	91.10	74.52	74.30	16.00	36.47	0.160
mEXNEX_{0}	89.72	89.26	70.92	70.69	11.19	35.71	0.112

Table A.1.4: Simulation results for Chapter 2: Operating characteristics for a simulation based on the realised sample size of 20, 10, 8, 18 and 7 across the 5 baskets for scenarios 1-6.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	6.36	4.83	10.47	6.14	7.39	69.49	0.305
BHM	11.15	9.11	8.46	11.07	8.59	70.66	0.293
CBHM	8.49	7.28	9.9	6.86	9.22	72.28	0.277
BMA	10.41	9.21	9.37	10.34	9.75	65.99	0.340
EXNEX	9.22	7.42	10.33	10.56	7.19	65.56	0.344
$\text{mEXNEX}_{1/13}$	9.48	12.30	10.31	11.70	8.04	58.98	0.410
mEXNEX_0	6.85	4.79	10.83	6.04	7.42	68.61	0.314
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	94.61	4.83	10.40	5.91	7.67	70.02	0.261
BHM	95.61	18.43	16.32	17.29	15.18	53.00	0.434
CBHM	94.60	7.75	12.19	7.72	10.26	68.58	0.272
BMA	96.05	13.74	13.66	14.82	13.16	59.76	0.373
EXNEX	95.28	12.00	10.77	11.68	8.80	61.92	0.346
$\text{mEXNEX}_{1/13}$	95.40	14.13	10.76	11.75	8.51	60.60	0.361
mEXNEX_0	94.27	4.93	10.45	5.85	7.47	69.97	0.258
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**		
Independent	94.14	73.33	10.92	5.57	7.83	54.14	0.222
BHM	97.62	87.94	21.31	21.56	22.44	47.25	0.431
CBHM	94.83	75.62	12.32	7.75	10.15	53.96	0.232
BMA	96.46	84.12	17.93	17.97	19.00	47.87	0.372
EXNEX	96.27	82.59	11.05	12.31	11.54	54.81	0.287
$\text{mEXNEX}_{1/13}$	95.90	84.97	11.49	12.36	11.09	57.33	0.290
mEXNEX_0	94.70	73.27	10.49	5.86	7.45	54.22	0.219
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**		
Independent	94.12	73.39	78.11	5.53	7.61	47.24	0.127
BHM	98.10	91.41	86.00	28.53	30.15	39.06	0.447
CBHM	95.20	75.65	79.04	9.27	10.93	45.85	0.157
BMA	97.09	88.34	82.69	21.92	25.58	37.73	0.410
EXNEX	96.70	87.15	78.40	13.43	17.31	47.47	0.267
$\text{mEXNEX}_{1/13}$	96.25	87.37	78.90	12.15	14.16	50.08	0.234
mEXNEX_0	94.70	73.44	77.75	5.89	7.30	47.11	0.128
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**		
Independent	94.92	73.29	78.10	90.83	7.96	45.46	0.080
BHM	98.72	94.08	90.92	98.39	51.36	35.06	0.514
CBHM	95.16	79.55	81.27	92.30	24.01	36.52	0.240
BMA	98.06	89.68	90.46	97.36	28.21	56.25	0.282
EXNEX	98.03	89.88	79.65	95.84	27.58	46.83	0.276
$\text{mEXNEX}_{1/13}$	96.27	88.98	80.34	95.95	17.93	53.30	0.179
mEXNEX_0	94.57	73.22	78.47	90.93	7.23	45.79	0.723
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	94.37	73.47	77.77	90.76	68.12	33.05	
BHM	99.39	96.97	94.72	99.01	94.80	87.29	
CBHM	95.70	84.24	84.82	94.06	80.99	59.07	
BMA	98.16	90.15	90.92	97.59	89.61	70.66	
EXNEX	98.29	89.75	83.66	96.37	88.21	64.59	
$\text{mEXNEX}_{1/13}$	96.77	89.98	79.90	95.60	78.18	52.27	
mEXNEX_0	94.57	73.11	77.92	90.90	69.20	33.54	

Table A.1.5: Simulation results for Chapter 2: Operating characteristics for a simulation based on the realised sample size of 20, 10, 8, 18 and 7 across the 5 baskets for scenarios 7-12.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	76.02	4.94	10.40	4.99	7.20	57.36	0.249
BHM	80.18	17.27	15.52	17.46	14.69	43.94	0.399
CBHM	75.84	9.30	13.08	8.98	11.39	53.72	0.267
BMA	80.10	13.81	13.52	15.06	13.48	48.14	0.371
EXNEX	78.75	12.00	10.56	11.91	8.77	50.38	0.341
$\text{mEXNEX}_{1/13}$	79.25	13.25	11.01	11.77	9.68	50.38	0.357
mEXNEX_0	76.02	4.94	10.40	5.90	7.20	56.84	0.257
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	76.02	47.90	57.64	4.86	7.20	19.18	0.118
BHM	88.00	74.06	68.46	27.85	27.75	22.18	0.414
CBHM	78.39	54.96	60.62	14.03	15.81	18.10	0.202
BMA	85.28	69.90	65.30	22.84	23.12	19.26	0.389
EXNEX	83.73	67.45	58.24	13.66	15.33	24.78	0.246
$\text{mEXNEX}_{1/13}$	81.89	66.64	59.23	12.33	14.58	23.85	0.238
mEXNEX_0	76.02	47.93	57.64	5.76	7.20	18.99	0.126
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	94.36	47.90	57.64	5.22	7.20	23.37	0.121
BHM	97.99	75.36	68.87	27.28	27.86	24.42	0.420
CBHM	95.02	52.26	60.02	10.73	12.31	22.56	0.171
BMA	97.25	69.49	64.70	21.19	23.47	22.24	0.379
EXNEX	96.76	68.53	58.20	13.04	15.93	28.11	0.249
$\text{mEXNEX}_{1/13}$	96.12	67.52	59.08	12.31	13.79	28.57	0.230
mEXNEX_0	94.36	47.90	57.64	5.86	7.20	23.21	0.127
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**		
Independent	94.36	72.77	57.64	67.87	7.20	24.85	0.072
BHM	98.87	93.39	79.56	90.52	46.96	28.35	0.496
CBHM	95.57	79.91	64.17	74.09	24.04	20.08	0.240
BMA	98.03	89.48	78.91	87.79	27.76	45.58	0.278
EXNEX	97.94	89.39	61.24	82.50	25.36	30.02	0.254
$\text{mEXNEX}_{1/13}$	96.66	88.21	61.64	82.01	18.18	33.81	0.182
mEXNEX_0	94.36	72.83	57.64	69.63	7.20	25.39	0.072
Scenario 11	**0.15**	**0.15**	**0.15**	**0.15**	**0.45**		
Independent	6.75	4.94	10.40	4.91	68.65	51.58	0.244
BHM	15.60	14.74	13.30	15.13	63.74	31.89	0.384
CBHM	8.47	7.22	11.26	6.76	68.78	49.76	0.248
BMA	12.07	10.87	10.99	12.22	70.21	45.60	0.328
EXNEX	10.71	9.97	10.36	11.60	68.89	44.60	0.335
$\text{mEXNEX}_{1/13}$	10.22	12.99	10.52	11.36	69.15	41.90	0.328
mEXNEX_0	6.75	4.94	10.40	5.90	68.65	51.06	0.253
Scenario 12	**0.15**	**0.15**	**0.45**	**0.15**	**0.45**		
Independent	6.75	4.94	78.46	5.19	68.65	45.88	0.160
BHM	19.81	19.69	77.71	19.10	72.94	30.79	0.416
CBHM	8.45	71.40	76.84	7.07	69.58	45.28	0.176
BMA	14.06	13.59	76.02	15.03	72.10	32.99	0.328
EXNEX	13.61	13.76	78.05	12.07	69.18	35.06	0.326
$\text{mEXNEX}_{1/13}$	12.05	14.42	78.53	11.57	69.70	35.86	0.319
mEXNEX_0	6.75	4.94	78.46	5.77	68.65	45.49	0.166

Table A.1.6: Simulation results for Chapter 2: Operating characteristics for a simulation based on the realised sample size of 20, 10, 8, 18 and 7 across the 5 baskets for scenarios 13-16.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 13	**0.15**	**0.45**	**0.45**	**0.15**	**0.45**		
Independent	6.75	72.77	78.46	4.87	68.65	35.11	0.113
BHM	24.43	88.84	83.42	24.96	80.68	33.21	0.393
CBHM	8.80	74.51	78.85	7.73	69.96	34.72	0.138
BMA	16.49	84.14	80.71	18.01	76.41	32.74	0.301
EXNEX	16.08	84.34	78.56	12.59	70.79	34.35	0.262
$\text{mEXNEX}_{1/13}$	12.96	85.24	78.71	11.93	70.96	36.75	0.230
mEXNEX_0	6.75	72.77	78.46	5.93	68.65	34.77	0.123
Scenario 14	**0.15**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	6.75	72.77	78.46	90.51	68.65	33.62	0.068
BHM	33.15	92.31	89.48	98.17	88.46	45.93	0.332
CBHM	10.60	75.58	79.79	91.72	71.10	33.01	0.106
BMA	16.88	89.36	86.45	97.33	87.04	56.35	0.169
EXNEX	16.72	89.33	79.41	96.53	79.03	45.97	0.167
$\text{mEXNEX}_{1/13}$	13.64	88.06	79.75	96.36	73.96	44.16	0.136
mEXNEX_0	6.75	72.82	78.46	91.20	68.65	33.72	0.068
Scenario 15	**0.45**	**0.15**	**0.15**	**0.15**	**0.45**		
Independent	94.36	4.94	10.40	5.05	68.65	52.35	0.191
BHM	97.24	22.52	20.75	21.14	76.40	38.11	0.433
CBHM	94.60	7.75	12.61	7.88	70.05	51.10	0.213
BMA	96.31	16.60	16.98	17.12	74.34	41.21	0.364
EXNEX	95.90	16.22	10.77	12.33	70.05	42.54	0.336
$\text{mEXNEX}_{1/13}$	95.77	15.27	11.14	11.85	70.29	43.45	0.325
mEXNEX_0	94.36	4.95	10.40	5.83	68.65	51.79	0.198
Scenario 16	**0.45**	**0.15**	**0.45**	**0.15**	**0.45**		
Independent	94.36	4.94	78.46	4.91	68.65	46.49	0.097
BHM	98.25	27.38	84.74	25.99	82.68	37.57	0.402
CBHM	94.71	8.36	79.27	8.70	70.39	44.99	0.134
BMA	97.20	17.75	82.49	20.40	79.90	38.51	0.338
EXNEX	97.15	17.80	78.74	12.98	72.42	38.90	0.278
$\text{mEXNEX}_{1/13}$	96.39	17.06	79.08	11.88	72.42	39.00	0.268
mEXNEX_0	94.36	4.94	78.46	6.00	68.65	45.88	0.107

A.2 Simulation Study for Realised Sample Size With Re-Calibrated Δ_α Values

In addition to those in the Chapter 2, a further simulation study was conducted on the realised sample size case of 20, 10, 8, 18 and 7 patients across the five baskets. In the previous study, the decision cut-off, Δ_α, was calibrated under a null scenario based on $n_k = 13$ patients in each basket and applied to the realised sample sizes. In this simulation, the Δ_α values are re-calibrated based on the unequal sample sizes to again achieve a basket specific type I error rate of 10% under the null scenario.

Table A.2.1: Re-calibrated Δ_α values for a simulation study comparing information borrowing methods based on realised sample sizes of 20, 10, 8, 18 and 7 across the five baskets as opposed to the planned sample size.

	$\Delta_{\alpha 1}$	$\Delta_{\alpha 2}$	$\Delta_{\alpha 3}$	$\Delta_{\alpha 4}$	$\Delta_{\alpha 5}$
Independent	0.859	0.862	0.922	0.899	0.786
BHM	0.852	0.811	0.806	0.844	0.806
CBHM	0.962	0.912	0.855	0.964	0.820
BMA	0.876	0.856	0.865	0.876	0.863
EXNEX	0.861	0.849	0.868	0.875	0.822
$\text{mEXNEX}_{1/13}$	0.877	0.887	0.907	0.903	0.836
mEXNEX_0	0.880	0.888	0.937	0.916	0.831

Table A.2.2: Operating characteristics for a simulation study to compare information borrowing models based on the realised sample size of 20, 10, 8, 18 and 7 across the baskets under data scenarios 1-6, with re-calibration of Δ_α to take into account the unequal sample sizes.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	10.45	10.75	9.81	9.38	9.39	59.19	0.408
BHM	9.34	10.17	10.18	9.62	10.48	69.84	0.316
CBHM	10.21	10.11	10.21	10.62	10.05	68.36	0.316
BMA	10.15	9.47	10.17	9.62	9.92	65.69	0.343
EXNEX	10.16	9.99	9.76	9.55	9.68	64.33	0.357
$\text{mEXNEX}_{1/13}$	10.14	9.12	9.98	9.52	9.54	62.45	0.376
mEXNEX_0	9.72	9.88	9.20	9.93	10.21	59.77	0.402
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	95.51	10.35	10.18	10.42	9.57	62.38	0.348
BHM	94.78	20.90	18.29	17.02	19.00	49.18	0.467
CBHM	94.92	13.37	12.04	13.12	9.97	60.71	35.23
BMA	95.62	13.87	13.23	13.57	13.62	60.06	0.366
EXNEX	95.52	15.93	10.54	11.64	14.32	57.48	0.392
$\text{mEXNEX}_{1/13}$	95.67	10.76	10.49	11.29	13.01	62.12	0.348
mEXNEX_0	95.22	10.53	8.94	9.51	9.86	63.11	0.398
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**		
Independent	95.75	80.2	10.07	9.91	9.35	56.38	0.266
BHM	96.58	89.87	24.72	19.78	25.81	44.34	0.468
CBHM	94.88	82.90	12.75	13.39	10.27	55.66	0.281
BMA	96.45	84.45	19.17	16.82	20.09	47.48	0.379
EXNEX	96.27	87.40	10.89	12.53	22.49	50.69	0.373
$\text{mEXNEX}_{1/13}$	96.36	81.36	10.74	11.47	15.13	51.41	0.315
mEXNEX_0	95.48	79.57	9.10	9.89	9.34	56.30	0.258
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**		
Independent	95.99	80.41	75.40	9.93	9.26	47.03	0.184
BHM	97.93	92.13	87.46	25.06	32.78	41.31	0.434
CBHM	95.66	84.73	78.77	14.29	11.27	49.50	0.205
BMA	97.23	89.25	82.89	19.95	25.72	39.84	0.395
EXNEX	97.41	89.41	78.13	12.79	26.96	43.29	0.356
$\text{mEXNEX}_{1/13}$	96.62	84.43	77.87	11.92	17.84	45.91	0.265
mEXNEX_0	95.36	79.89	74.86	9.58	10.11	46.68	0.189
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**		
Independent	95.65	79.53	75.31	94.53	9.14	48.92	0.091
BHM	98.77	94.74	93.09	98.14	56.04	32.36	0.560
CBHM	95.80	86.36	80.95	95.93	24.02	43.59	0.240
BMA	98.11	89.33	91.39	97.80	27.96	56.46	0.280
EXNEX	98.11	90.17	80.77	95.75	28.48	47.97	0.285
$\text{mEXNEX}_{1/13}$	97.47	86.96	77.82	96.32	19.48	51.70	0.195
mEXNEX_0	95.56	79.96	75.49	93.87	9.98	49.22	0.098
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	95.75	79.81	75.37	94.22	71.10	38.88	
BHM	99.11	97.17	96.10	98.86	95.87	89.11	
CBHM	96.18	91.23	85.48	96.36	82.04	64.41	
BMA	97.79	90.15	92.17	97.53	89.91	71.44	
EXNEX	98.10	90.30	83.96	96.19	89.55	65.24	
$\text{mEXNEX}_{1/13}$	97.50	87.58	78.17	95.73	80.25	50.10	
mEXNEX_0	95.82	79.31	74.10	93.79	70.47	37.04	

Table A.2.3: Operating characteristics for a simulation study to compare information borrowing models based on the realised sample size of 20, 10, 8, 18 and 7 across the baskets under data scenarios 7-12, with re-calibration of Δ_α to take into account the unequal sample sizes.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	78.83	8.21	9.13	8.93	7.84	55.53	0.301
BHM	78.00	19.50	17.96	16.31	17.63	40.12	0.428
CBHM	60.63	10.06	16.72	6.66	15.74	42.67	0.297
BMA	79.81	14.83	14.57	14.46	14.35	47.13	0.382
EXNEX	79.02	13.78	8.98	10.96	7.50	49.89	0.337
$\text{mEXNEX}_{1/13}$	80.24	11.12	10.11	10.42	13.19	51.91	0.342
mEXNEX_0	79.68	10.70	9.35	9.65	9.67	52.75	0.340
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	78.90	54.35	53.30	9.02	8.06	19.57	0.165
BHM	86.15	76.36	71.75	25.51	31.72	23.43	0.421
CBHM	62.79	55.41	62.90	10.38	20.11	14.05	0.222
BMA	84.68	71.07	66.84	21.77	24.01	20.14	0.389
EXNEX	84.09	69.48	55.88	11.64	10.13	28.66	0.196
$\text{mEXNEX}_{1/13}$	82.72	63.68	57.25	11.49	19.66	21.76	0.280
mEXNEX_0	79.86	59.62	54.05	9.56	9.77	21.98	0.184
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	95.38	54.25	53.27	9.16	7.85	23.28	0.163
BHM	97.38	77.04	72.24	25.19	31.62	26.40	0.426
CBHM	88.33	52.07	61.30	8.06	15.46	20.90	0.179
BMA	97.04	70.87	66.15	20.11	24.20	23.53	0.378
EXNEX	96.96	70.47	56.85	11.73	10.24	32.29	0.197
$\text{mEXNEX}_{1/13}$	96.43	63.35	57.48	11.54	17.70	26.03	0.261
mEXNEX_0	95.28	57.62	54.19	9.42	9.70	24.48	0.182
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**		
Independent	95.24	77.09	53.15	76.13	7.90	27.07	0.079
BHM	98.66	93.97	82.81	89.51	51.73	25.62	0.517
CBHM	89.02	78.64	64.80	67.77	27.64	15.17	0.276
BMA	98.01	89.56	79.79	86.92	27.83	45.37	0.278
EXNEX	98.01	89.60	57.61	81.70	20.38	30.05	0.204
$\text{mEXNEX}_{1/13}$	97.07	86.30	58.17	81.43	21.49	31.68	0.215
mEXNEX_0	95.42	79.57	53.87	77.13	9.61	28.49	0.096
Scenario 11	**0.15**	**0.15**	**0.15**	**0.15**	**0.45**		
Independent	9.17	8.21	8.99	8.97	69.47	47.74	0.309
BHM	13.40	16.85	15.25	13.73	67.61	34.84	0.388
CBHM	4.43	8.21	14.34	5.60	72.15	53.70	0.232
BMA	11.72	12.01	11.74	11.83	70.62	44.75	0.336
EXNEX	10.96	11.43	7.49	9.76	68.84	46.29	0.301
$\text{mEXNEX}_{1/13}$	10.99	10.20	10.10	9.99	70.76	44.71	0.336
mEXNEX_0	9.74	9.60	9.23	9.40	71.35	47.65	0.326
Scenario 12	**0.15**	**0.15**	**0.45**	**0.15**	**0.45**		
Independent	9.18	8.21	74.48	9.15	69.43	39.71	0.243
BHM	17.76	21.71	79.62	17.42	76.08	33.29	0.404
CBHM	3.92	7.67	79.50	5.28	71.84	49.32	0.131
BMA	13.61	14.94	77.06	14.58	72.83	33.63	0.337
EXNEX	14.02	15.26	69.33	11.24	68.75	29.17	0.338
$\text{mEXNEX}_{1/13}$	12.97	12.04	77.20	10.84	72.08	35.89	0.299
mEXNEX_0	10.20	10.19	74.94	9.72	71.25	39.35	0.268

Table A.2.4: Operating characteristics for a simulation study to compare information borrowing models based on the realised sample size of 20, 10, 8, 18 and 7 across the baskets under data scenarios 13-16, with re-calibration of Δ_α to take into account the unequal sample sizes.

	% Reject					% All Correct	FWER
Sample Size	20	10	8	18	7		
Scenario 13	**0.15**	**0.45**	**0.45**	**0.15**	**0.45**		
Independent	9.14	76.96	74.06	9.09	69.41	33.02	0.175
BHM	21.90	89.88	85.67	23.05	84.44	39.16	0.362
CBHM	4.29	74.62	79.71	5.77	71.35	37.84	0.080
BMA	16.20	85.69	81.50	16.87	77.37	35.37	0.289
EXNEX	16.27	86.11	76.40	11.70	69.08	35.04	0.261
$\text{mEXNEX}_{1/13}$	14.06	81.32	77.90	11.55	74.14	34.28	0.239
mEXNEX_0	10.31	79.30	75.32	9.46	71.20	34.64	0.189
Scenario 14	**0.15**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	9.26	77.06	74.50	93.93	69.29	34.41	0.093
BHM	31.41	93.18	91.31	97.98	89.91	48.62	0.314
CBHM	6.85	75.15	80.19	89.48	72.20	34.61	0.069
BMA	16.75	89.55	87.46	97.15	87.68	57.31	0.168
EXNEX	16.73	89.50	78.43	96.35	71.87	40.29	0.167
$\text{mEXNEX}_{1/13}$	14.73	85.30	78.49	96.17	77.37	42.06	0.147
mEXNEX_0	9.80	78.84	75.07	94.37	71.28	35.87	0.098
Scenario 15	**0.45**	**0.15**	**0.15**	**0.15**	**0.45**		
Independent	95.18	8.24	9.07	9.29	69.31	49.74	0.243
BHM	96.52	24.35	23.96	19.89	79.27	37.98	0.451
CBHM	87.52	7.87	13.46	5.75	71.08	48.10	0.256
BMA	96.23	17.26	17.89	16.21	75.09	41.05	0.371
EXNEX	96.06	17.08	10.26	11.73	69.01	42.27	0.342
$\text{mEXNEX}_{1/13}$	96.13	12.46	10.38	11.20	73.12	46.82	0.294
mEXNEX_0	95.48	9.86	9.38	9.68	71.24	49.98	0.263
Scenario 16	**0.45**	**0.15**	**0.45**	**0.15**	**0.45**		
Independent	95.31	8.03	74.65	8.94	69.44	41.92	0.163
BHM	97.69	30.04	86.54	24.47	85.75	39.04	0.417
CBHM	87.95	8.95	79.80	6.54	71.74	43.14	0.117
BMA	97.06	18.03	83.23	19.08	81.00	40.45	0.329
EXNEX	97.16	17.94	77.80	11.81	69.59	38.52	0.276
$\text{mEXNEX}_{1/13}$	96.83	14.61	78.28	11.65	76.06	42.30	0.242
mEXNEX_0	95.47	9.97	75.02	9.58	71.18	42.06	0.186

Figure A.2.1: Percentage of rejections of the null hypothesis for each information borrowing method under data scenarios 1-10, based on the realised sample sizes of 20, 10, 8, 18 and 7, with re-calibration of Δ_α to take into account unequal sample sizes.

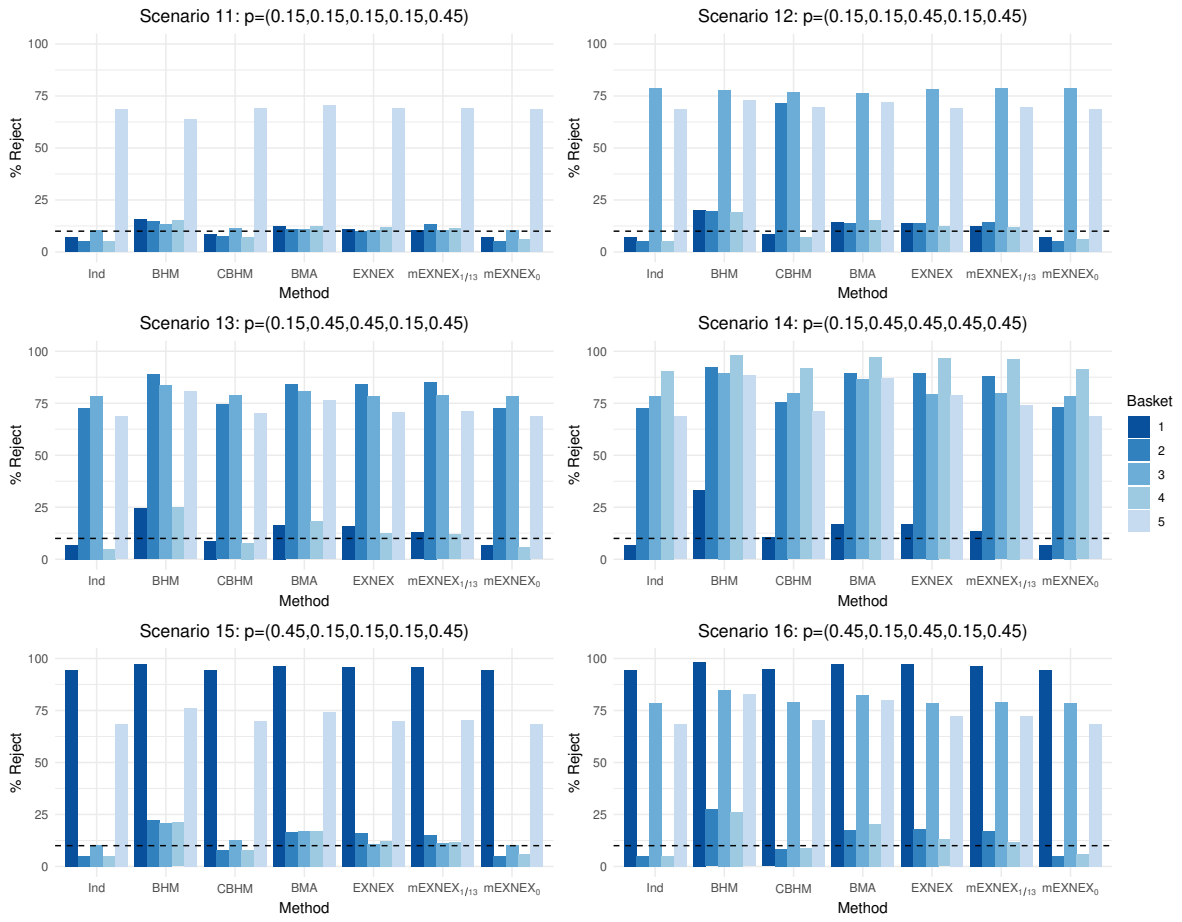


Figure A.2.2: Percentage of rejections of the null hypothesis for each information borrowing method under data scenarios 11-16, based on the realised sample sizes of 20, 10, 8, 18 and 7, with re-calibration of Δ_α to take into account unequal sample sizes.

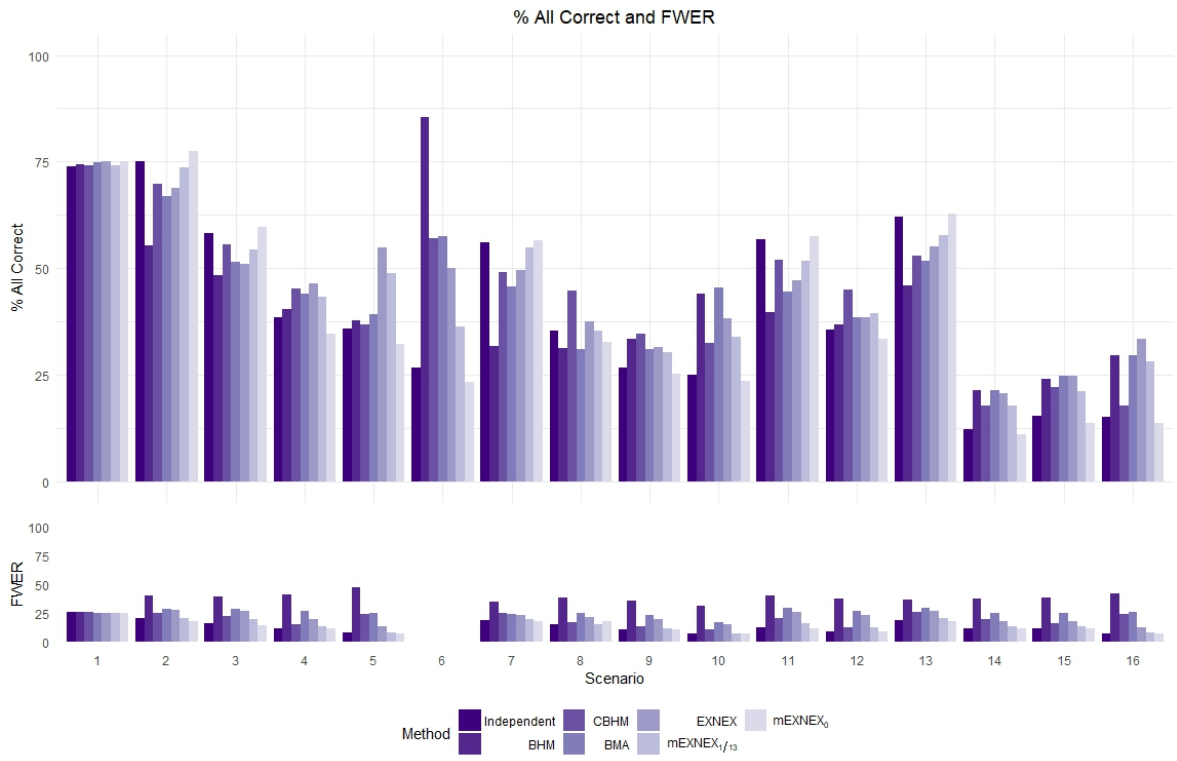


Figure A.2.3: The family-wise error rate (FWER) and percentage of times correct inference is made across all baskets (% All Correct) for each information borrowing method under each data scenario based on realised sample sizes of 20, 10, 8, 18 and 7, with re-calibration of Δ_α to take into account unequal sample sizes.

A.3 Estimation Ability of Information Borrowing Models

Provided in Tables A.3.1, A.3.2 (planned sample size scenario), A.3.3, A.3.4, A.3.5 (realised sample size scenario), A.3.6, A.3.7 and A.3.8 (realised sample size scenario with re-calibrated Δ_α) are the mean posterior point estimates of p_k across the 10,000 simulations in the simulation studies presented in 2. In brackets are the standard deviation of these mean estimates.

Table A.3.1: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on a planned sample size of 13 per basket under scenarios 1-6.

	Mean Point Estimate (standard deviation)				
Sample Size	13	13	13	13	13
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.151 (0.097)	0.150 (0.096)	0.151 (0.098)	0.151 (0.098)	0.153 (0.098)
BHM	0.150 (0.067)	0.149 (0.066)	0.149 (0.068)	0.150 (0.067)	0.151 (0.068)
CBHM	0.149 (0.084)	0.149 (0.085)	0.149 (0.084)	0.151 (0.086)	0.149 (0.085)
BMA	0.165 (0.074)	0.164 (0.073)	0.164 (0.074)	0.163 (0.074)	0.164 (0.075)
EXNEX	0.160 (0.079)	0.159 (0.079)	0.159 (0.080)	0.157 (0.078)	0.159 (0.089)
$\text{mEXNEX}_{1/13}$	0.157 (0.083)	0.157 (0.082)	0.156 (0.083)	0.156 (0.082)	0.154 (0.081)
mEXNEX_0	0.160 (0.090)	0.160 (0.089)	0.160 (0.089)	0.161 (0.090)	0.159 (0.079)
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.459 (0.137)	0.150 (0.097)	0.151 (0.098)	0.151 (0.098)	0.151 (0.097)
BHM	0.381 (0.132)	0.167 (0.077)	0.167 (0.078)	0.169 (0.078)	0.167 (0.078)
CBHM	0.441 (0.147)	0.152 (0.095)	0.154 (0.096)	0.154 (0.096)	0.150 (0.095)
BMA	0.418 (0.137)	0.172 (0.079)	0.174 (0.079)	0.174 (0.079)	0.172 (0.078)
EXNEX	0.422 (0.132)	0.164 (0.081)	0.166 (0.082)	0.166 (0.082)	0.164 (0.082)
$\text{mEXNEX}_{1/13}$	0.432 (0.135)	0.158 (0.084)	0.156 (0.083)	0.158 (0.084)	0.157 (0.084)
mEXNEX_0	0.442 (0.132)	0.161 (0.091)	0.160 (0.090)	0.161 (0.090)	0.161 (0.089)
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**
Independent	0.446 (0.137)	0.452 (0.138)	0.150 (0.096)	0.151 (0.098)	0.151 (0.098)
BHM	0.404 (0.124)	0.402 (0.124)	0.180 (0.084)	0.182 (0.084)	0.182 (0.084)
CBHM	0.445 (0.142)	0.445 (0.142)	0.154 (0.099)	0.152 (0.098)	0.154 (0.099)
BMA	0.418 (0.127)	0.420 (0.129)	0.185 (0.087)	0.186 (0.088)	0.185 (0.087)
EXNEX	0.425 (0.127)	0.424 (0.129)	0.175 (0.086)	0.175 (0.085)	0.173 (0.086)
$\text{mEXNEX}_{1/13}$	0.432 (0.134)	0.430 (0.134)	0.168 (0.088)	0.166 (0.088)	0.167 (0.088)
mEXNEX_0	0.443 (0.130)	0.431 (0.129)	0.163 (0.089)	0.165 (0.089)	0.164 (0.090)
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**
Independent	0.449 (0.138)	0.449 (0.138)	0.448 (0.138)	0.152 (0.098)	0.153 (0.098)
BHM	0.415 (0.117)	0.416 (0.116)	0.416 (0.117)	0.200 (0.090)	0.203 (0.088)
CBHM	0.447 (0.139)	0.445 (0.137)	0.444 (0.137)	0.155 (0.101)	0.157 (0.102)
BMA	0.431 (0.117)	0.428 (0.116)	0.429 (0.117)	0.196 (0.096)	0.197 (0.096)
EXNEX	0.433 (0.123)	0.431 (0.120)	0.430 (0.123)	0.183 (0.091)	0.184 (0.090)
$\text{mEXNEX}_{1/13}$	0.437 (0.130)	0.437 (0.129)	0.436 (0.130)	0.177 (0.092)	0.176 (0.090)
mEXNEX_0	0.443 (0.128)	0.442 (0.130)	0.444 (0.130)	0.171 (0.089)	0.170 (0.088)
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**
Independent	0.449 (0.137)	0.448 (0.138)	0.449 (0.138)	0.450 (0.137)	0.152 (0.098)
BHM	0.428 (0.106)	0.427 (0.107)	0.427 (0.107)	0.427 (0.106)	0.238 (0.097)
CBHM	0.446 (0.135)	0.445 (0.135)	0.445 (0.135)	0.447 (0.137)	0.164 (0.112)
BMA	0.440 (0.108)	0.443 (0.108)	0.443 (0.108)	0.441 (0.107)	0.203 (0.102)
EXNEX	0.439 (0.116)	0.440 (0.116)	0.439 (0.117)	0.441 (0.114)	0.193 (0.095)
$\text{mEXNEX}_{1/13}$	0.444 (0.124)	0.443 (0.124)	0.444 (0.124)	0.446 (0.126)	0.180 (0.092)
mEXNEX_0	0.444 (0.129)	0.446 (0.130)	0.446 (0.128)	0.446 (0.129)	0.173 (0.088)
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**
Independent	0.447 (0.138)	0.450 (0.138)	0.448 (0.138)	0.450 (0.137)	0.450 (0.139)
BHM	0.451 (0.088)	0.450 (0.089)	0.451 (0.091)	0.449 (0.089)	0.449 (0.089)
CBHM	0.449 (0.124)	0.450 (0.123)	0.451 (0.122)	0.450 (0.123)	0.450 (0.123)
BMA	0.450 (0.104)	0.449 (0.103)	0.449 (0.100)	0.450 (0.104)	0.448 (0.103)
EXNEX	0.449 (0.107)	0.450 (0.109)	0.448 (0.110)	0.450 (0.110)	0.449 (0.109)
$\text{mEXNEX}_{1/13}$	0.448 (0.118)	0.448 (0.120)	0.450 (0.120)	0.447 (0.119)	0.446 (0.119)
mEXNEX_0	0.444 (0.127)	0.447 (0.128)	0.444 (0.127)	0.448 (0.126)	0.448 (0.127)

Table A.3.2: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on a planned sample size of 13 per basket under scenarios 7-10.

	Mean Point Estimate (standard deviation)				
Sample Size	13	13	13	13	13
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.350 (0.131)	0.159 (0.098)	0.151 (0.096)	0.151 (0.096)	0.151 (0.097)
BHM	0.294 (0.112)	0.164 (0.072)	0.164 (0.072)	0.164 (0.071)	0.164 (0.072)
CBHM	0.335 (0.138)	0.153 (0.092)	0.154 (0.091)	0.154 (0.090)	0.154 (0.092)
BMA	0.324 (0.119)	0.171 (0.077)	0.172 (0.076)	0.172 (0.076)	0.171 (0.076)
EXNEX	0.329 (0.120)	0.164 (0.081)	0.165 (0.080)	0.165 (0.079)	0.165 (0.080)
$\text{mEXNEX}_{1/13}$	0.337 (0.129)	0.150 (0.083)	0.159 (0.082)	0.159 (0.082)	0.159 (0.083)
mEXNEX_0	0.348 (0.126)	0.161 (0.090)	0.161 (0.089)	0.162 (0.088)	0.161 (0.090)
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.350 (0.131)	0.348 (0.132)	0.350 (0.131)	0.151 (0.096)	0.151 (0.097)
BHM	0.320 (0.102)	0.319 (0.103)	0.320 (0.101)	0.195 (0.078)	0.195 (0.079)
CBHM	0.344 (0.130)	0.342 (0.131)	0.344 (0.129)	0.160 (0.098)	0.160 (0.099)
BMA	0.337 (0.106)	0.335 (0.107)	0.336 (0.106)	0.192 (0.085)	0.192 (0.086)
EXNEX	0.338 (0.113)	0.336 (0.114)	0.337 (0.113)	0.181 (0.084)	0.180 (0.084)
$\text{mEXNEX}_{1/13}$	0.340 (0.121)	0.338 (0.122)	0.340 (0.121)	0.176 (0.085)	0.175 (0.086)
mEXNEX_0	0.348 (0.123)	0.346 (0.124)	0.348 (0.123)	0.169 (0.086)	0.169 (0.087)
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.449 (0.137)	0.348 (0.132)	0.350 (0.131)	0.151 (0.096)	0.151 (0.097)
BHM	0.399 (0.117)	0.328 (0.104)	0.329 (0.103)	0.197 (0.082)	0.197 (0.083)
CBHM	0.443 (0.141)	0.345 (0.131)	0.347 (0.129)	0.158 (0.099)	0.157 (0.100)
BMA	0.420 (0.121)	0.340 (0.108)	0.342 (0.107)	0.194 (0.088)	0.193 (0.089)
EXNEX	0.425 (0.124)	0.340 (0.114)	0.341 (0.113)	0.182 (0.086)	0.181 (0.086)
$\text{mEXNEX}_{1/13}$	0.434 (0.132)	0.340 (0.122)	0.342 (0.121)	0.176 (0.087)	0.175 (0.088)
mEXNEX_0	0.442 (0.130)	0.347 (0.124)	0.349 (0.123)	0.169 (0.087)	0.169 (0.088)
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**
Independent	0.449 (0.137)	0.448 (0.138)	0.350 (0.131)	0.350 (0.130)	0.151 (0.097)
BHM	0.413 (0.107)	0.412 (0.108)	0.348 (0.098)	0.349 (0.096)	0.227 (0.090)
CBHM	0.444 (0.137)	0.442 (0.138)	0.349 (0.096)	0.350 (0.126)	0.164 (0.109)
BMA	0.429 (0.111)	0.438 (0.111)	0.359 (0.103)	0.359 (0.101)	0.200 (0.096)
EXNEX	0.432 (0.117)	0.430 (0.118)	0.352 (0.110)	0.353 (0.108)	0.191 (0.091)
$\text{mEXNEX}_{1/13}$	0.438 (0.126)	0.437 (0.127)	0.353 (0.117)	0.353 (0.116)	0.180 (0.091)
mEXNEX_0	0.444 (0.129)	0.442 (0.130)	0.352 (0.122)	0.352 (0.121)	0.172 (0.087)

Table A.3.3: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes for scenarios 1-6.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.151 (0.078)	0.154 (0.110)	0.152 (0.123)	0.151 (0.083)	0.153 (0.131)
BHM	0.149 (0.061)	0.150 (0.074)	0.151 (0.077)	0.149 (0.064)	0.151 (0.079)
CBHM	0.148 (0.071)	0.152 (0.094)	0.153 (0.101)	0.149 (0.075)	0.151 (0.079)
BMA	0.160 (0.065)	0.167 (0.082)	0.171 (0.088)	0.161 (0.067)	0.174 (0.094)
EXNEX	0.156 (0.068)	0.161 (0.087)	0.166 (0.096)	0.157 (0.072)	0.167 (0.098)
$\text{mEXNEX}_{1/13}$	0.153 (0.072)	0.160 (0.096)	0.166 (0.106)	0.153 (0.076)	0.168 (0.112)
mEXNEX_0	0.162 (0.074)	0.169 (0.103)	0.178 (0.112)	0.165 (0.078)	0.183 (0.120)
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.449 (0.111)	0.151 (0.111)	0.154 (0.124)	0.151 (0.083)	0.152 (0.131)
BHM	0.408 (0.110)	0.172 (0.090)	0.173 (0.096)	0.165 (0.075)	0.175 (0.099)
CBHM	0.445 (0.119)	0.153 (0.110)	0.153 (0.120)	0.153 (0.084)	0.155 (0.129)
BMA	0.427 (0.112)	0.178 (0.090)	0.183 (0.096)	0.170 (0.075)	0.185 (0.098)
EXNEX	0.432 (0.109)	0.170 (0.096)	0.173 (0.101)	0.164 (0.077)	0.177 (0.105)
$\text{mEXNEX}_{1/13}$	0.440 (0.109)	0.163 (0.100)	0.167 (0.109)	0.158 (0.079)	0.171 (0.116)
mEXNEX_0	0.446 (0.106)	0.172 (0.102)	0.179 (0.115)	0.164 (0.078)	0.177 (0.105)
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**
Independent	0.448 (0.111)	0.450 (0.157)	0.154 (0.124)	0.151 (0.082)	0.155 (0.132)
BHM	0.415 (0.105)	0.403 (0.134)	0.190 (0.099)	0.178 (0.079)	0.194 (0.103)
CBHM	0.446 (0.114)	0.443 (0.159)	0.155 (0.124)	0.151 (0.084)	0.157 (0.131)
BMA	0.429 (0.107)	0.418 (0.137)	0.198 (0.105)	0.179 (0.081)	0.204 (0.107)
EXNEX	0.434 (0.107)	0.422 (0.140)	0.184 (0.105)	0.169 (0.078)	0.188 (0.109)
$\text{mEXNEX}_{1/13}$	0.443 (0.108)	0.437 (0.145)	0.175 (0.111)	0.163 (0.079)	0.179 (0.116)
mEXNEX_0	0.447 (0.106)	0.445 (0.142)	0.180 (0.110)	0.167 (0.078)	0.186 (0.116)
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**
Independent	0.450 (0.110)	0.451 (0.156)	0.447 (0.176)	0.149 (0.083)	0.155 (0.129)
BHM	0.424 (0.101)	0.414 (0.128)	0.410 (0.137)	0.192 (0.084)	0.218 (0.104)
CBHM	0.447 (0.113)	0.445 (0.157)	0.443 (0.176)	0.155 (0.088)	0.157 (0.132)
BMA	0.433 (0.101)	0.428 (0.127)	0.425 (0.137)	0.185 (0.085)	0.217 (0.113)
EXNEX	0.435 (0.105)	0.430 (0.135)	0.426 (0.145)	0.176 (0.081)	0.201 (0.110)
$\text{mEXNEX}_{1/13}$	0.444 (0.109)	0.442 (0.145)	0.440 (0.159)	0.165 (0.078)	0.190 (0.114)
mEXNEX_0	0.446 (0.107)	0.443 (0.142)	0.443 (0.154)	0.166 (0.077)	0.191 (0.109)
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**
Independent	0.448 (0.112)	0.448 (0.157)	0.448 (0.175)	0.449 (0.117)	0.153 (0.131)
BHM	0.436 (0.091)	0.432 (0.111)	0.429 (0.117)	0.435 (0.094)	0.277 (0.106)
CBHM	0.445 (0.108)	0.446 (0.151)	0.444 (0.167)	0.444 (0.114)	0.187 (0.152)
BMA	0.443 (0.091)	0.443 (0.114)	0.443 (0.124)	0.446 (0.095)	0.226 (0.123)
EXNEX	0.445 (0.098)	0.440 (0.127)	0.437 (0.135)	0.445 (0.101)	0.219 (0.116)
$\text{mEXNEX}_{1/13}$	0.448 (0.106)	0.446 (0.139)	0.445 (0.153)	0.448 (0.109)	0.198 (0.113)
mEXNEX_0	0.445 (0.106)	0.444 (0.143)	0.441 (0.154)	0.445 (0.111)	0.190 (0.106)
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**
Independent	0.450 (0.110)	0.449 (0.158)	0.449 (0.175)	0.450 (0.116)	0.447 (0.184)
BHM	0.451 (0.083)	0.449 (0.099)	0.449 (0.103)	0.450 (0.086)	0.451 (0.105)
CBHM	0.450 (0.104)	0.448 (0.138)	0.448 (0.154)	0.451 (0.107)	0.451 (0.161)
BMA	0.450 (0.090)	0.451 (0.113)	0.449 (0.124)	0.449 (0.091)	0.448 (0.127)
EXNEX	0.451 (0.094)	0.449 (0.121)	0.445 (0.129)	0.449 (0.098)	0.448 (0.134)
$\text{mEXNEX}_{1/13}$	0.446 (0.104)	0.448 (0.138)	0.446 (0.149)	0.449 (0.091)	0.448 (0.127)
mEXNEX_0	0.447 (0.106)	0.443 (0.141)	0.444 (0.154)	0.445 (0.112)	0.442 (0.161)

Table A.3.4: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes of 20, 10, 8, 18 and 7 patients across the 5 baskets for scenarios 7-12.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.350 (0.106)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.154 (0.130)
BHM	0.312 (0.098)	0.168 (0.083)	0.171 (0.086)	0.165 (0.069)	0.172 (0.090)
CBHM	0.339 (0.113)	0.155 (0.105)	0.158 (0.113)	0.154 (0.080)	0.160 (0.121)
BMA	0.330 (0.100)	0.178 (0.087)	0.183 (0.090)	0.172 (0.072)	0.185 (0.095)
EXNEX	0.336 (0.102)	0.169 (0.092)	0.173 (0.097)	0.163 (0.073)	0.176 (0.102)
$\text{mEXNEX}_{1/13}$	0.344 (0.106)	0.163 (0.098)	0.167 (0.106)	0.158 (0.076)	0.171 (0.113)
mEXNEX_0	0.351 (0.102)	0.173 (0.103)	0.178 (0.112)	0.165 (0.076)	0.181 (0.120)
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.350 (0.106)	0.347 (0.150)	0.349 (0.165)	0.151 (0.082)	0.154 (0.130)
BHM	0.326 (0.091)	0.138 (0.112)	0.316 (0.118)	0.188 (0.074)	0.206 (0.089)
CBHM	0.344 (0.107)	0.341 (0.147)	0.342 (0.160)	0.158 (0.086)	0.167 (0.126)
BMA	0.339 (0.094)	0.336 (0.117)	0.337 (0.125)	0.185 (0.078)	0.206 (0.101)
EXNEX	0.340 (0.097)	0.336 (0.125)	0.336 (0.134)	0.174 (0.075)	0.194 (0.102)
$\text{mEXNEX}_{1/13}$	0.346 (0.103)	0.344 (0.137)	0.345 (0.148)	0.166 (0.076)	0.186 (0.109)
mEXNEX_0	0.351 (0.102)	0.351 (0.136)	0.354 (0.146)	0.166 (0.075)	0.190 (0.109)
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.450 (0.111)	0.347 (0.150)	0.349 (0.165)	0.151 (0.082)	0.154 (0.130)
BHM	0.413 (0.103)	0.329 (0.115)	0.328 (0.122)	0.189 (0.078)	0.209 (0.097)
CBHM	0.445 (0.115)	0.349 (0.148)	0.347 (0.162)	0.155 (0.086)	0.161 (0.131)
BMA	0.428 (0.104)	0.344 (0.119)	0.345 (0.126)	0.185 (0.081)	0.210 (0.106)
EXNEX	0.433 (0.105)	0.341 (0.127)	0.342 (0.135)	0.174 (0.077)	0.196 (0.106)
$\text{mEXNEX}_{1/13}$	0.442 (0.109)	0.345 (0.138)	0.347 (0.149)	0.165 (0.077)	0.186 (0.110)
mEXNEX_0	0.446 (0.106)	0.352 (0.136)	0.354 (0.146)	0.166 (0.075)	0.190 (0.109)
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**
Independent	0.450 (0.111)	0.446 (0.157)	0.349 (0.165)	0.350 (0.110)	0.154 (0.130)
BHM	0.423 (0.092)	0.414 (0.113)	0.355 (0.111)	0.355 (0.087)	0.258 (0.100)
CBHM	0.443 (0.111)	0.438 (0.152)	0.351 (0.155)	0.350 (0.106)	0.185 (0.142)
BMA	0.434 (0.093)	0.429 (0.118)	0.365 (0.119)	0.363 (0.090)	0.223 (0.115)
EXNEX	0.437 (0.099)	0.429 (0.128)	0.356 (0.129)	0.355 (0.097)	0.212 (0.110)
$\text{mEXNEX}_{1/13}$	0.444 (0.106)	0.439 (0.142)	0.356 (0.144)	0.353 (0.104)	0.194 (0.110)
mEXNEX_0	0.446 (0.106)	0.442 (0.142)	0.355 (0.146)	0.352 (0.104)	0.191 (0.108)
Scenario 11	**0.15**	**0.15**	**0.15**	**0.15**	**0.45**
Independent	0.150 (0.079)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.449 (0.186)
BHM	0.160 (0.066)	0.165 (0.081)	0.168 (0.084)	0.161 (0.067)	0.352 (0.161)
CBHM	0.152 (0.077)	0.154 (0.105)	0.157 (0.114)	0.152 (0.079)	0.429 (0.201)
BMA	0.165 (0.066)	0.172 (0.083)	0.177 (0.087)	0.166 (0.068)	0.406 (0.170)
EXNEX	0.160 (0.069)	0.167 (0.089)	0.171 (0.094)	0.161 (0.071)	0.408 (0.166)
$\text{mEXNEX}_{1/13}$	0.155 (0.073)	0.162 (0.096)	0.167 (0.103)	0.156 (0.075)	0.432 (0.170)
mEXNEX_0	0.162 (0.074)	0.172 (0.101)	0.181 (0.107)	0.166 (0.076)	0.440 (0.161)
Scenario 12	**0.15**	**0.15**	**0.45**	**0.15**	**0.45**
Independent	0.150 (0.079)	0.151 (0.110)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.170 (0.069)	0.179 (0.086)	0.381 (0.149)	0.171 (0.071)	0.378 (0.156)
CBHM	0.152 (0.079)	0.154 (0.110)	0.442 (0.179)	0.153 (0.082)	0.442 (0.192)
BMA	0.172 (0.070)	0.182 (0.088)	0.410 (0.156)	0.174 (0.072)	0.408 (0.164)
EXNEX	0.165 (0.071)	0.175 (0.091)	0.415 (0.154)	0.166 (0.075)	0.441 (0.161)
$\text{mEXNEX}_{1/13}$	0.159 (0.074)	0.169 (0.097)	0.431 (0.160)	0.160 (0.076)	0.430 (0.168)
mEXNEX_0	0.162 (0.074)	0.175 (0.097)	0.441 (0.152)	0.166 (0.075)	0.441 (0.161)

Table A.3.5: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes of 20, 10, 8, 18 and 7 patients across the 5 baskets for scenarios 13-16.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 13	**0.15**	**0.45**	**0.45**	**0.15**	**0.15**
Independent	0.150 (0.079)	0.446 (0.157)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.181 (0.073)	0.402 (0.134)	0.400 (0.142)	0.184 (0.075)	0.397 (0.149)
CBHM	0.152 (0.080)	0.444 (0.159)	0.447 (0.175)	0.153 (0.084)	0.446 (0.188)
BMA	0.179 (0.076)	0.421 (0.139)	0.420 (0.147)	0.182 (0.078)	0.419 (0.154)
EXNEX	0.171 (0.074)	0.424 (0.140)	0.422 (0.149)	0.173 (0.076)	0.421 (0.157)
$\text{mEXNEX}_{1/13}$	0.163 (0.074)	0.435 (0.147)	0.436 (0.158)	0.165 (0.076)	0.434 (0.167)
mEXNEX_0	0.164 (0.072)	0.441 (0.143)	0.442 (0.152)	0.166 (0.075)	0.441 (0.160)
Scenario 14	**0.15**	**0.45**	**0.45**	**0.45**	**0.45**
Independent	0.150 (0.079)	0.446 (0.157)	0.449 (0.173)	0.450 (0.115)	0.449 (0.186)
BHM	0.206 (0.083)	0.421 (0.122)	0.420 (0.129)	0.429 (0.100)	0.419 (0.135)
CBHM	0.155 (0.086)	0.445 (0.156)	0.448 (0.172)	0.448 (0.116)	0.448 (0.185)
BMA	0.185 (0.084)	0.437 (0.123)	0.438 (0.130)	0.442 (0.101)	0.438 (0.137)
EXNEX	0.178 (0.078)	0.435 (0.131)	0.435 (0.140)	0.441 (0.105)	0.435 (0.147)
$\text{mEXNEX}_{1/13}$	0.166 (0.075)	0.442 (0.143)	0.444 (0.153)	0.446 (0.110)	0.443 (0.162)
mEXNEX_0	0.164 (0.072)	0.441 (0.143)	0.443 (0.152)	0.446 (0.109)	0.441 (0.160)
Scenario 15	**0.45**	**0.15**	**0.15**	**0.15**	**0.45**
Independent	0.450 (0.111)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.149 (0.186)
BHM	0.415 (0.107)	0.184 (0.093)	0.189 (0.097)	0.176 (0.076)	0.392 (0.153)
CBHM	0.447 (0.115)	0.153 (0.112)	0.155 (0.122)	0.153 (0.083)	0.449 (0.189)
BMA	0.428 (0.109)	0.189 (0.095)	0.196 (0.099)	0.178 (0.078)	0.412 (0.156)
EXNEX	0.433 (0.107)	0.177 (0.096)	0.183 (0.102)	0.169 (0.076)	0.417 (0.159)
$\text{mEXNEX}_{1/13}$	0.442 (0.109)	0.168 (0.100)	0.175 (0.107)	0.161 (0.077)	0.431 (0.168)
mEXNEX_0	0.446 (0.106)	0.175 (0.100)	0.182 (0.107)	0.161 (0.077)	0.431 (0.158)
Scenario 16	**0.45**	**0.15**	**0.45**	**0.15**	**0.45**
Independent	0.450 (0.111)	0.151 (0.110)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.422 (0.102)	0.202 (0.095)	0.407 (0.139)	0.189 (0.080)	0.405 (0.146)
CBHM	0.448 (0.113)	0.155 (0.114)	0.447 (0.174)	0.154 (0.085)	0.447 (0.187)
BMA	0.434 (0.104)	0.201 (0.102)	0.424 (0.138)	0.185 (0.083)	0.423 (0.146)
EXNEX	0.436 (0.105)	0.188 (0.098)	0.426 (0.146)	0.175 (0.078)	0.425 (0.154)
$\text{mEXNEX}_{1/13}$	0.443 (0.108)	0.177 (0.099)	0.437 (0.156)	0.166 (0.077)	0.436 (0.165)
mEXNEX_0	0.446 (0.106)	0.178 (0.099)	0.437 (0.156)	0.166 (0.077)	0.436 (0.165)

Table A.3.6: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes of 20, 10, 8, 18 and 7 patients with re-calibration of Δ_α under scenarios 1-6.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.150 (0.079)	0.153 (0.111)	0.153 (0.122)	0.150 (0.084)	0.153 (0.132)
BHM	0.149 (0.062)	0.149 (0.074)	0.151 (0.078)	0.149 (0.064)	0.151 (0.079)
CBHM	0.148 (0.071)	0.150 (0.092)	0.152 (0.099)	0.147 (0.074)	0.156 (0.107)
BMA	0.158 (0.065)	0.165 (0.081)	0.171 (0.087)	0.161 (0.068)	0.173 (0.091)
EXNEX	0.156 (0.069)	0.162 (0.089)	0.165 (0.096)	0.156 (0.071)	0.169 (0.098)
$\text{mEXNEX}_{1/13}$	0.153 (0.073)	0.160 (0.096)	0.165 (0.105)	0.155 (0.076)	0.168 (0.112)
mEXNEX_0	0.162 (0.075)	0.170 (0.104)	0.176 (0.112)	0.165 (0.078)	0.182 (0.121)
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.448 (0.110)	0.152 (0.109)	0.153 (0.122)	0.150 (0.083)	0.151 (0.129)
BHM	0.409 (0.112)	0.171 (0.091)	0.172 (0.095)	0.165 (0.074)	0.175 (0.099)
CBHM	0.445 (0.118)	0.154 (0.110)	0.154 (0.119)	0.152 (0.084)	0.156 (0.129)
BMA	0.426 (0.112)	0.177 (0.089)	0.183 (0.095)	0.171 (0.075)	0.187 (0.098)
EXNEX	0.432 (0.110)	)0.172 (0.095)	0.175 (0.101)	0.163 (0.076)	0.176 (0.104)
$\text{mEXNEX}_{1/13}$	0.444 (0.111)	0.161 (0.097)	0.168 (0.111)	0.160 (0.079)	0.169 (0.113)
mEXNEX_0	0.446 (0.106)	0.175 (0.103)	0.179 (0.114)	0.164 (0.078)	0.181 (0.119)
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**
Independent	0.452 (0.113)	0.448 (0.157)	0.152 (0.123)	0.150 (0.083)	0.152 (0.128)
BHM	0.417 (0.106)	0.401 (0.136)	0.192 (0.100)	0.177 (0.078)	0.196 (0.103)
CBHM	0.448 (0.116)	0.446 (0.160)	0.156 (0.125)	0.152 (0.085)	0.156 (0.133)
BMA	0.427 (0.106)	0.416 (0.136)	0.201 (0.104)	0.181 (0.081)	0.205 (0.106)
EXNEX	0.434 (0.107)	0.425 (0.140)	0.186 (0.104)	0.169 (0.078)	0.189 (0.108)
$\text{mEXNEX}_{1/13}$	0.443 (0.107)	0.436 (0.145)	0.176 (0.109)	0.162 (0.079)	0.178 (0.115)
mEXNEX_0	0.448 (0.106)	0.440 (0.140)	0.181 (0.110)	0.167 (0.078)	0.186 (0.117)
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**
Independent	0.449 (0.112)	0.446 (0.157)	0.449 (0.176)	0.152 (0.084)	0.156 (0.132)
BHM	0.424 (0.102)	0.412 (0.129)	0.410 (0.139)	0.192 (0.084)	0.216 (0.104)
CBHM	0.448 (0.114)	0.447 (0.157)	0.445 (0.175)	0.154 (0.087)	0.159 (0.134)
BMA	0.435 (0.102)	0.431 (0.128)	0.428 (0.136)	0.187 (0.087)	0.215 (0.113)
EXNEX	0.438 (0.104)	0.429 (0.136)	0.428 (0.146)	0.176 (0.081)	0.200 (0.110)
$\text{mEXNEX}_{1/13}$	0.444 (0.109)	0.436 (0.144)	0.439 (0.157)	0.165 (0.077)	0.190 (0.112)
mEXNEX_0	0.445 (0.106)	0.445 (0.143)	0.444 (0.156)	0.165 (0.077)	0.192 (0.108)
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**
Independent	0.449 (0.110)	0.449 (0.156)	0.450 (0.173)	0.448 (0.117)	0.154 (0.130)
BHM	0.435 (0.090)	0.432 (0.112)	0.431 (0.119)	0.435 (0.092)	0.278 (0.106)
CBHM	0.446 (0.109)	0.445 (0.151)	0.443 (0.168)	0.445 (0.114)	0.187 (0.151)
BMA	0.445 (0.090)	0.444 (0.114)	0.443 (0.124)	0.444 (0.094)	0.227 (0.123)
EXNEX	0.443 (0.099)	0.439 (0.128)	0.437 (0.138)	0.442 (0.102)	0.218 (0.115)
$\text{mEXNEX}_{1/13}$	0.447 (0.106)	0.447 (0.140)	0.444 (0.151)	0.444 (0.111)	0.197 (0.112)
mEXNEX_0	0.446 (0.107)	0.445 (0.142)	0.441 (0.155)	0.446 (0.112)	0.192 (0.107)
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**
Independent	0.451 (0.110)	0.445 (0.158)	0.450 (0.175)	0.449 (0.118)	0.450 (0.187)
BHM	0.449 (0.082)	0.449 (0.097)	0.449 (0.101)	0.451 (0.085)	0.450 (0.104)
CBHM	0.449 (0.104)	0.449 (0.138)	0.450 (0.151)	0.449 (0.107)	0.452 (0.159)
BMA	0.451 (0.090)	0.451 (0.113)	0.448 (0.122)	0.449 (0.092)	0.447 (0.130)
EXNEX	0.449 (0.094)	0.445 (0.120)	0.449 (0.130)	0.447 (0.099)	0.443 (0.135)
$\text{mEXNEX}_{1/13}$	0.448 (0.104)	0.446 (0.137)	0.446 (0.148)	0.449 (0.108)	0.445 (0.158)
mEXNEX_0	0.447 (0.105)	0.440 (0.142)	0.443 (0.153)	0.446 (0.110)	0.435 (0.159)

Table A.3.7: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes of 20, 10, 8, 18 and 7 patients across the 5 baskets with re-calibration of Δ_α under scenarios 7-12.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 7	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**
Independent	0.350 (0.106)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.154 (0.130)
BHM	0.312 (0.098)	0.168 (0.083)	0.171 (0.086)	0.165 (0.069)	0.172 (0.090)
CBHM	0.337 (0.116)	0.153 (0.106)	0.155 (0.115)	0.152 (0.080)	0.157 (0.123)
BMA	0.330 (0.100)	0.178 (0.087)	0.183 (0.090)	0.172 (0.072)	0.185 (0.095)
EXNEX	0.336 (0.102)	0.169 (0.092)	0.173 (0.097)	0.163 (0.073)	0.176 (0.102)
$\text{mEXNEX}_{1/13}$	0.344 (0.106)	0.163 (0.098)	0.167 (0.106)	0.158 (0.076)	0.171 (0.113)
mEXNEX_0	0.351 (0.102)	0.173 (0.103)	0.178 (0.112)	0.165 (0.076)	0.181 (0.120)
Scenario 8	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.350 (0.106)	0.347 (0.150)	0.349 (0.165)	0.151 (0.082)	0.154 (0.130)
BHM	0.326 (0.091)	0.138 (0.112)	0.316 (0.118)	0.188 (0.074)	0.206 (0.089)
CBHM	0.341 (0.111)	0.338 (0.150)	0.340 (0.164)	0.156 (0.085)	0.162 (0.128)
BMA	0.339 (0.094)	0.336 (0.117)	0.337 (0.125)	0.185 (0.078)	0.206 (0.101)
EXNEX	0.340 (0.097)	0.336 (0.125)	0.336 (0.134)	0.174 (0.075)	0.194 (0.102)
$\text{mEXNEX}_{1/13}$	0.346 (0.103)	0.344 (0.137)	0.345 (0.148)	0.166 (0.076)	0.186 (0.109)
mEXNEX_0	0.351 (0.102)	0.351 (0.136)	0.354 (0.146)	0.166 (0.075)	0.190 (0.109)
Scenario 9	**0.45**	**0.35**	**0.35**	**0.15**	**0.15**
Independent	0.450 (0.111)	0.347 (0.150)	0.349 (0.165)	0.151 (0.082)	0.154 (0.130)
BHM	0.413 (0.103)	0.329 (0.115)	0.328 (0.122)	0.189 (0.078)	0.209 (0.097)
CBHM	0.443 (0.118)	0.344 (0.149)	0.346 (0.164)	0.154 (0.086)	0.157 (0.133)
BMA	0.428 (0.104)	0.344 (0.119)	0.345 (0.126)	0.185 (0.081)	0.210 (0.106)
EXNEX	0.433 (0.105)	0.341 (0.127)	0.342 (0.135)	0.174 (0.077)	0.196 (0.106)
$\text{mEXNEX}_{1/13}$	0.442 (0.109)	0.345 (0.138)	0.347 (0.149)	0.165 (0.077)	0.186 (0.110)
mEXNEX_0	0.446 (0.106)	0.352 (0.136)	0.354 (0.146)	0.166 (0.075)	0.190 (0.109)
Scenario 10	**0.45**	**0.45**	**0.35**	**0.35**	**0.15**
Independent	0.450 (0.111)	0.446 (0.157)	0.349 (0.165)	0.350 (0.110)	0.154 (0.130)
BHM	0.423 (0.092)	0.414 (0.113)	0.355 (0.111)	0.355 (0.087)	0.258 (0.100)
CBHM	0.434 (0.121)	0.431 (0.160)	0.343 (0.160)	0.343 (0.111)	0.174 (0.142)
BMA	0.434 (0.093)	0.429 (0.118)	0.365 (0.119)	0.363 (0.090)	0.223 (0.115)
EXNEX	0.437 (0.099)	0.429 (0.128)	0.356 (0.129)	0.355 (0.097)	0.212 (0.110)
$\text{mEXNEX}_{1/13}$	0.444 (0.106)	0.439 (0.142)	0.356 (0.144)	0.353 (0.104)	0.194 (0.110)
mEXNEX_0	0.446 (0.106)	0.442 (0.142)	0.355 (0.146)	0.352 (0.104)	0.191 (0.108)
Scenario 11	**0.15**	**0.15**	**0.15**	**0.15**	**0.45**
Independent	0.150 (0.079)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.449 (0.186)
BHM	0.160 (0.066)	0.165 (0.081)	0.168 (0.084)	0.161 (0.067)	0.352 (0.161)
CBHM	0.151 (0.077)	0.152 (0.106)	0.155 (0.116)	0.151 (0.080)	0.430 (0.203)
BMA	0.165 (0.066)	0.172 (0.083)	0.177 (0.087)	0.166 (0.068)	0.406 (0.170)
EXNEX	0.160 (0.069)	0.167 (0.089)	0.171 (0.094)	0.161 (0.071)	0.408 (0.166)
$\text{mEXNEX}_{1/13}$	0.155 (0.073)	0.162 (0.096)	0.167 (0.103)	0.156 (0.075)	0.432 (0.170)
mEXNEX_0	0.162 (0.074)	0.172 (0.101)	0.181 (0.107)	0.166 (0.076)	0.440 (0.161)
Scenario 12	**0.15**	**0.15**	**0.45**	**0.15**	**0.45**
Independent	0.150 (0.079)	0.151 (0.110)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.170 (0.069)	0.179 (0.086)	0.381 (0.149)	0.171 (0.071)	0.378 (0.156)
CBHM	0.151 (0.079)	0.152 (0.111)	0.442 (0.180)	0.152 (0.082)	0.442 (0.193)
BMA	0.172 (0.070)	0.182 (0.088)	0.410 (0.156)	0.174 (0.072)	0.408 (0.164)
EXNEX	0.165 (0.071)	0.175 (0.091)	0.415 (0.154)	0.166 (0.075)	0.441 (0.161)
$\text{mEXNEX}_{1/13}$	0.159 (0.074)	0.169 (0.097)	0.431 (0.160)	0.160 (0.076)	0.430 (0.168)
mEXNEX_0	0.162 (0.074)	0.175 (0.097)	0.441 (0.152)	0.166 (0.075)	0.441 (0.161)

Table A.3.8: Simulation results for Chapter 2: Mean point estimates of p_k across the simulations (standard deviations) based on realised sample sizes of 20, 10, 8, 18 and 7 patients across the 5 baskets with re-calibration of Δ_α under scenarios 13-16.

	Mean Point Estimate (standard deviation)				
Sample Size	20	10	8	18	7
Scenario 13	**0.15**	**0.45**	**0.45**	**0.15**	**0.15**
Independent	0.150 (0.079)	0.446 (0.157)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.181 (0.073)	0.402 (0.134)	0.400 (0.142)	0.184 (0.075)	0.397 (0.149)
CBHM	0.152 (0.081)	0.444 (0.159)	0.447 (0.175)	0.153 (0.084)	0.446 (0.188)
BMA	0.179 (0.076)	0.421 (0.139)	0.420 (0.147)	0.182 (0.078)	0.419 (0.154)
EXNEX	0.171 (0.074)	0.424 (0.140)	0.422 (0.149)	0.173 (0.076)	0.421 (0.157)
$\text{mEXNEX}_{1/13}$	0.163 (0.074)	0.435 (0.147)	0.436 (0.158)	0.165 (0.076)	0.434 (0.167)
mEXNEX_0	0.164 (0.072)	0.441 (0.143)	0.442 (0.152)	0.166 (0.075)	0.441 (0.160)
Scenario 14	**0.15**	**0.45**	**0.45**	**0.45**	**0.45**
Independent	0.150 (0.079)	0.446 (0.157)	0.449 (0.173)	0.450 (0.115)	0.449 (0.186)
BHM	0.206 (0.083)	0.421 (0.122)	0.420 (0.129)	0.429 (0.100)	0.419 (0.135)
CBHM	0.154 (0.086)	0.444 (0.158)	0.446 (0.174)	0.446 (0.119)	0.447 (0.187)
BMA	0.185 (0.084)	0.437 (0.123)	0.438 (0.130)	0.442 (0.101)	0.438 (0.137)
EXNEX	0.178 (0.078)	0.435 (0.131)	0.435 (0.140)	0.441 (0.105)	0.435 (0.147)
$\text{mEXNEX}_{1/13}$	0.166 (0.075)	0.442 (0.143)	0.444 (0.153)	0.446 (0.110)	0.443 (0.162)
mEXNEX_0	0.164 (0.072)	0.441 (0.143)	0.443 (0.152)	0.446 (0.109)	0.441 (0.160)
Scenario 15	**0.45**	**0.15**	**0.15**	**0.15**	**0.45**
Independent	0.450 (0.111)	0.151 (0.111)	0.153 (0.121)	0.151 (0.082)	0.149 (0.186)
BHM	0.415 (0.107)	0.184 (0.093)	0.189 (0.097)	0.176 (0.076)	0.392 (0.153)
CBHM	0.446 (0.116)	0.151 (0.113)	0.153 (0.123)	0.152 (0.083)	0.445 (0.190)
BMA	0.428 (0.109)	0.189 (0.095)	0.196 (0.099)	0.178 (0.078)	0.412 (0.156)
EXNEX	0.433 (0.107)	0.177 (0.096)	0.183 (0.102)	0.169 (0.076)	0.417 (0.159)
$\text{mEXNEX}_{1/13}$	0.442 (0.109)	0.168 (0.100)	0.175 (0.107)	0.161 (0.077)	0.431 (0.168)
mEXNEX_0	0.446 (0.106)	0.175 (0.100)	0.182 (0.107)	0.161 (0.077)	0.431 (0.158)
Scenario 16	**0.45**	**0.15**	**0.45**	**0.15**	**0.45**
Independent	0.450 (0.111)	0.151 (0.110)	0.449 (0.173)	0.151 (0.082)	0.449 (0.186)
BHM	0.422 (0.102)	0.202 (0.095)	0.407 (0.139)	0.189 (0.080)	0.405 (0.146)
CBHM	0.447 (0.114)	0.153 (0.114)	0.447 (0.175)	0.153 (0.085)	0.447 (0.188)
BMA	0.434 (0.104)	0.201 (0.102)	0.424 (0.138)	0.185 (0.083)	0.423 (0.146)
EXNEX	0.436 (0.105)	0.188 (0.098)	0.426 (0.146)	0.175 (0.078)	0.425 (0.154)
$\text{mEXNEX}_{1/13}$	0.443 (0.108)	0.177 (0.099)	0.437 (0.156)	0.166 (0.077)	0.436 (0.165)
mEXNEX_0	0.446 (0.106)	0.178 (0.099)	0.437 (0.156)	0.166 (0.077)	0.436 (0.165)

A.4 Simulation Study in Which the Truth Vector, p, is Varied

There are an infinite number of data scenarios one could fall in when conducting clinical trial analysis, the results presented in the previously mentioned simulation studies in Chapter 2 are only a subset of these feasible possible data scenarios. The data scenarios used in the simulation studies were selected to cover a wide range of cases, however, some important cases may have not been investigated.

To overcome this, a further simulation study was conducted within which, rather than fixing the true probability of success parameter prior to the study, for every simulation run a new random truth vector, p, was generated and data simulated from a Binomial distribution using these p values. In order to ensure equal chances of lying in the null and non-null case, p was selected with uniform probability across the ranges [0,0.15] and [0.35,0.5].

A total of 20,000 simulations for each borrowing method was run under the three simulation cases: planned sample size of $n_k = 13$ in each basket, realised sample sizes of 20, 10, 8, 18 and 7 without re-calibration of Δ_α and the realised sample size case with re-calibration. For each method and setting, we find the following operating characteristics:

- Type I error rate - the percentage of times the null was rejected out of the cases where the null was in fact true. This is computed for each basket.
- Power - the percentage of times the null was rejected out of the cases where the true response rate was non-null. This is also computed for each basket.
- Percentage of all correct inference (% Correct) - the percentage of times the correct decision regarding whether to accept/reject the null was made across all 5 baskets.
- Family-wise error rate (FWER) - the percentage of times at least one type I error

was made across the 5 baskets (excluding the global alternative cases where p_k is non-null in all k baskets.

A.4.1 Planned Sample Size

Figure A.4.1 presents results for the planned sample size case in which 13 patients were observed in each of the 5 baskets. The top section of the figure demonstrates the type I error rate under each of the 7 methods. Similar to results presented in the planned sample size simulation in Chapter 2, the BHM and BMA have the highest error rates at approximately 5.0% and 4.51% respectively. All methods have errors less than or equal to the nominal 10% level. The reduced error rates comes from, in some cases, the true response rate lying well below the null level of 15%. The cut-off value Δ_α was calibrated for each method under a null scenario where the true response rate is 0.15. When a basket has a true response rate less than 0.15, the Δ_α value becomes conservative as it is easier to correctly identify that the treatment is ineffective. Under each of the borrowing models there is some degree of pull towards the common mean, which is most evident in the BHM and BMA case, and thus all have a higher error rate than under an independent analysis. Both the standard EXNEX and mEXNEX$_{1/13}$ model have almost identical rates both at around 3.2% each, whilst the mEXNEX$_0$ has a lower error rate of 2.7%.

When considering power, those methods with higher error rates also demonstrate the greatest power. The BHM has a power value increased 6% to that of the independent model but that came with the inflation of error as mentioned above. Again both the EXNEX model and mEXNEX$_{1/13}$ model perform almost identically with a power of 86%. The mEXNEX$_0$ model has lower power at 83.9% but this is still a 2% improvement over the independent model. The CBHM has very similar power and error rate to the independent analysis.

The percentage of times correct conclusion was made across all 5 baskets is presented

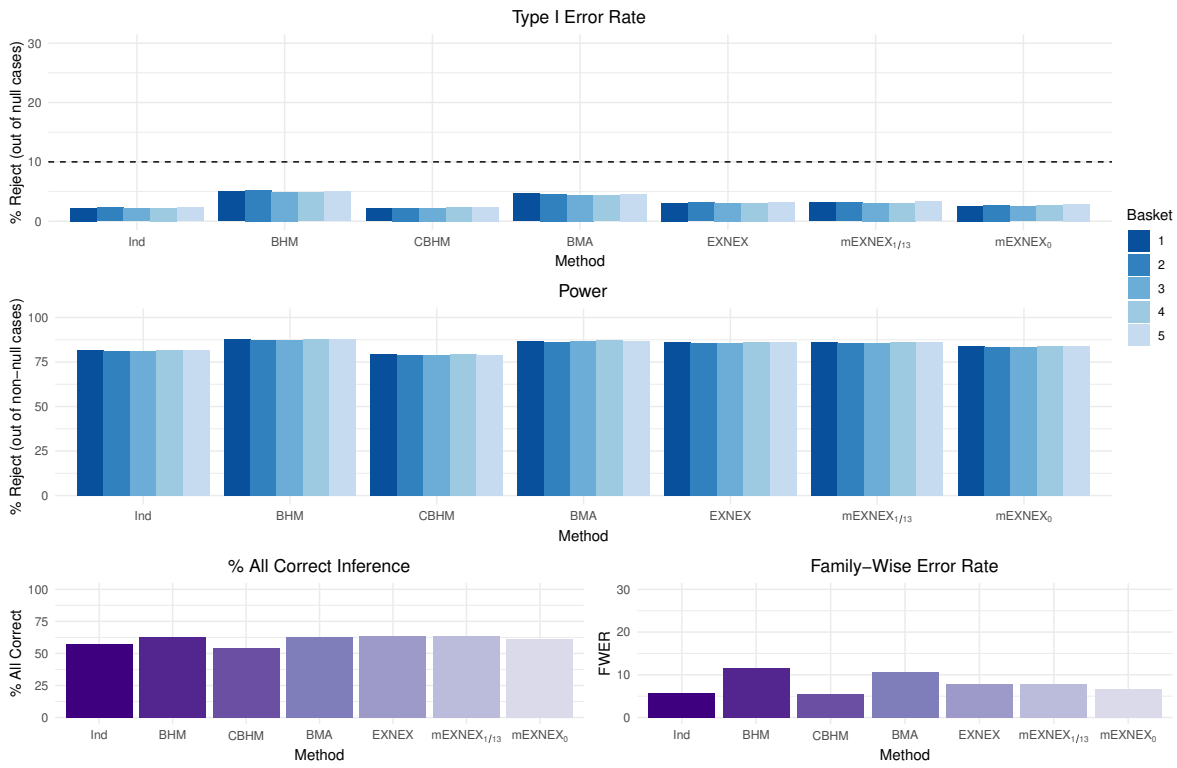


Figure A.4.1: Operating characteristics under varied truths for the planned sample size case presented in Chapter 2 where 13 patients are observed in each basket.

in the bottom left of Figure A.4.1. The BMA approach alongside the BHM, EXNEX and mEXNEX$_{1/13}$ model all make the correct conclusion across baskets in 63% of the simulations. However, the modified EXNEX approach with $c = 0$ makes slightly fewer all correction conclusions at 61.2% but this is still greater than an independent analysis which has a value of 57.6%.

A more substantial difference in methods is observed when looking at the family-wise error rate. Methods that demonstrated lower type I error rate also present lower FWERs, with the independent analysis giving the lowest error alongside the CBHM. This must be weighed up with the lower percentage of all correct inference and power that these two methods possess. The BHM and BMA have much larger FWER values, as expected based on the inflated type I error rate.

To summarise, in the planned sample size case when the true response rate is varied, the BHM and BMA continue to display undesirable error rates whilst the independent

analysis and CBHM lack power. The modified EXNEX model with $c = 1/13$ performs almost identically to the standard EXNEX model, whereas, when a more conservative cut-off value $c = 0$ is implemented, error rates are reduced by 0.5% from the standard EXNEX model but with a 2.1% reduction in power (but still a 2.4% improvement over an independent analysis).

A.4.2 Realised Sample Size

In the realised sample size case, basket sample sizes are equal to 20, 10, 8, 18 and 7 and Δ_α is calibrated based on the planned sample size of $n_k = 13$ in each of the k baskets. Results are presented in Figure A.4.2. Similar error rates are observed as in the planned sample size case, with the BHM and BMA approach having inflated error rates with higher errors in baskets where the sample size is small. In this case the mEXNEX$_0$ model performs almost identically to an independent analysis due to the discreetness of the data making it impossible for a basket to not be analysed as independent in first step of the mEXNEX$_c$ procedure.

The EXNEX and mEXNEX$_{1/13}$ again behave similarly in all metrics, thus little would be gained by using the modified EXNEX approach in this case (particularly for the choice of c made). The mEXNEX$_{1/13}$ does generate a 0.3% higher probability of all correct inference across the baskets compared to the EXNEX model whilst giving the same FWER.

Looking at the percentage of all correct inference across the 5 baskets and the family-wise error rates on a whole, performance of methods are very similar to the planned sample size case but with uniformly lower values for the first metric. The exception is the mEXNEX$_0$ model, which now is identical to the independent approach for the aforementioned reasons.

Overall, we conclude from these results that again a cut-off of $c = 0$ is not appropriate in the realised sample size case, however, performance when $c = 1/13$ is selected

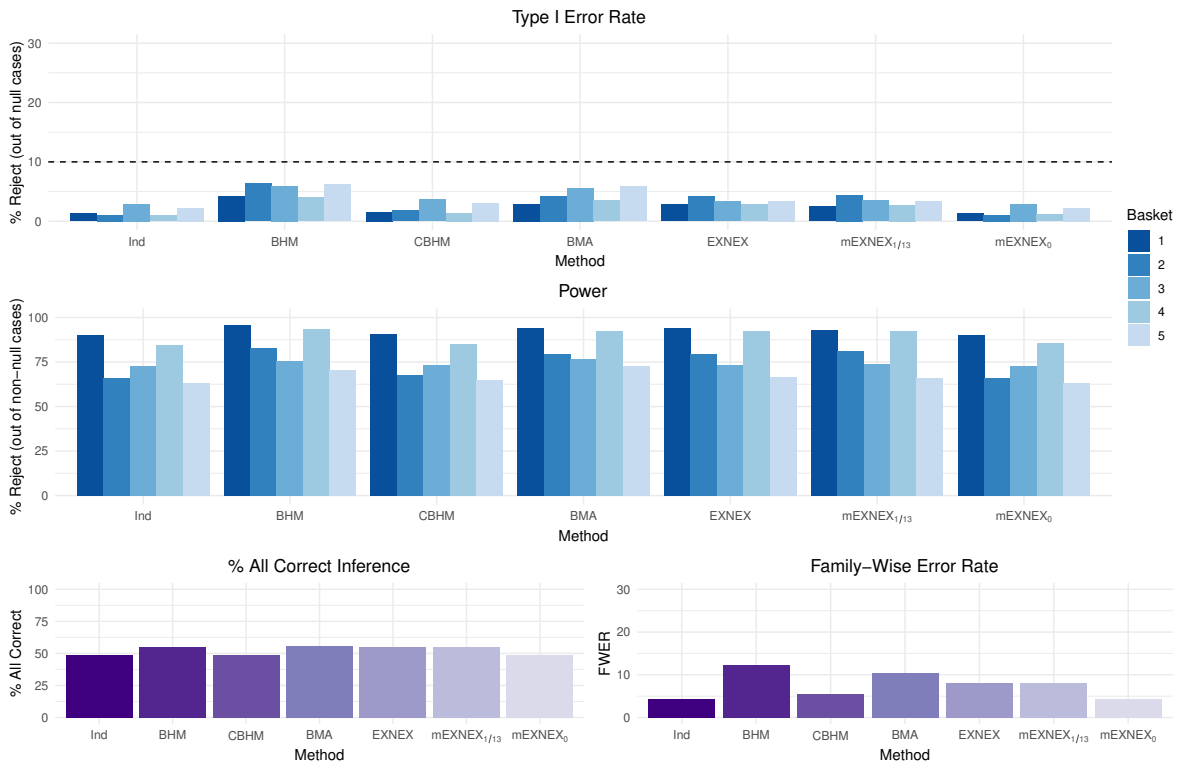


Figure A.4.2: Operating characteristics under varied truths for the realised sample size case presented in Chapter 2 where 20, 10, 8, 18 and 7 patients are observed in the 5 baskets and Δ_α is not re-calibrated

produces very similar results to the standard EXNEX model. Other values of c could be beneficial here but this was calibrated based on the planned sample size case. Power for smaller basket sizes tend to be considerably lower, so methods that borrow more strongly show clear benefits in power improvement. For example, the BHM improves power over an independent analysis by 7.2% in basket 5 when the sample size is just 7. The mEXNEX$_{1/13}$ model also demonstrates power improvement over an independent analysis by 2.8%.

A.4.3 Realised Sample Size with Re-Calibrated Δ_α

The above took the decision cut-off value Δ_α calibrated based on the planned sample size of 13 patients in each basket applied to the realised sample sizes of 20, 10, 8, 18 and 7 for the 5 baskets. In this section these Δ_α values are re-calibrated based on the

realised sample sizes. Results are akin to the realised sample size without re-calibration case and hence discussion is omitted.

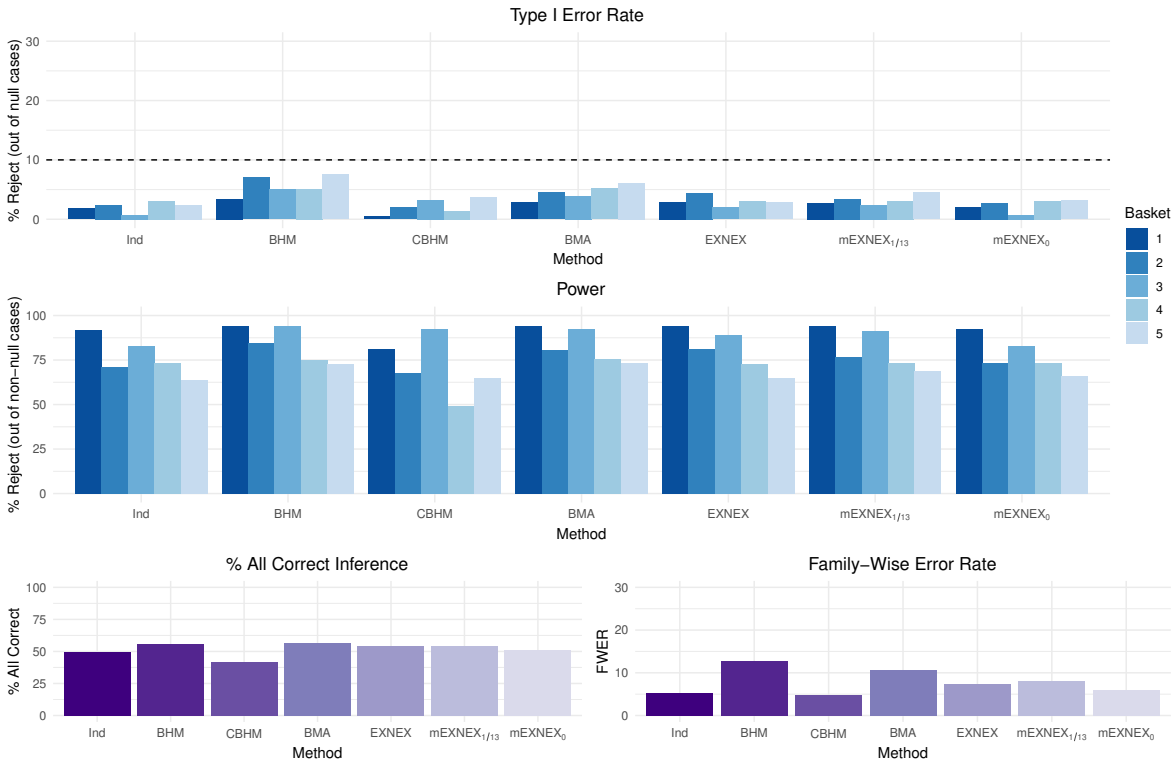


Figure A.4.3: Operating characteristics under varied truths for the realised sample size case presented in Chapter 2 where 20, 10, 8, 18 and 7 patients are observed in the 5 baskets and Δ_α is re-calibrated

A.5 Evaluation of the 1-step mEXNEX$_c$ Models Compared to the Proposed 2-step mEXNEX$_c$ Model

The specification of the mEXNEX$_c$ model outlined in Section 2.2.6 of Chapter 2, requires a two step procedure:

- Step 1: Remove clearly heterogeneous baskets to be analysed independently, setting $\pi_k = 0$ in the EXNEX model. This is conducted based on some pre-defined cut-off value c which is compared to the minimum pair-wise difference in responses.

- Step 2: Of remaining baskets, compute pairwise Hellinger distances and set the

prior borrowing probability π_k, to be the average of these distances (excluding the distance to itself).

Here we explore why the use of both of these steps is advantageous compared to making just one of these alterations to the standard EXNEX model. We consider four model settings:

1. The standard EXNEX model where $\pi_k = 0.5$ for all K baskets.
2. The EXNEX model with just step 2 i.e. no removal of heterogeneous baskets but Hellinger distances used to define the π_k values. Denote this as $\text{EXNEX}_{\text{Hell}}$.
3. The EXNEX model with just step 1, i.e. removing heterogeneous baskets and assigning remaining baskets a borrowing probability of 0.5. Denote this as EXNEXR_c.
4. The mEXNEX_c model as outlined in Section 2.2.6 which implements both steps.

A.5.1 Simulation Study Based on the Motivating VE-BASKET trial

To make such a comparison we initially consider the simulation setting outlined in Section 2.3 of Chapter 2 which is based on the motivating VE-BASKET trial. Results presented in Figure A.5.1 are based on a planned and equal sample size of $n_k = 13$ patients in each of the k baskets and a cut-off of $c = 0$ implemented for the methods that require removal of heterogeneous baskets. Figure A.5.1 displays the percentage of simulated data sets in which the null hypothesis is rejected for each basket and model setting under the data scenarios outlined in Table 2.3.1 in Chapter 2. When the null is true, the bars represent the basket's type I error rate, else it is the power.

From Figure A.5.1 one can clearly see that when heterogeneous baskets are not removed from the borrowing component (i.e. in the EXNEX and $\text{EXNEX}_{\text{Hell}}$ models) an inflation in error rate is evident. Whereas, models that take this removal step

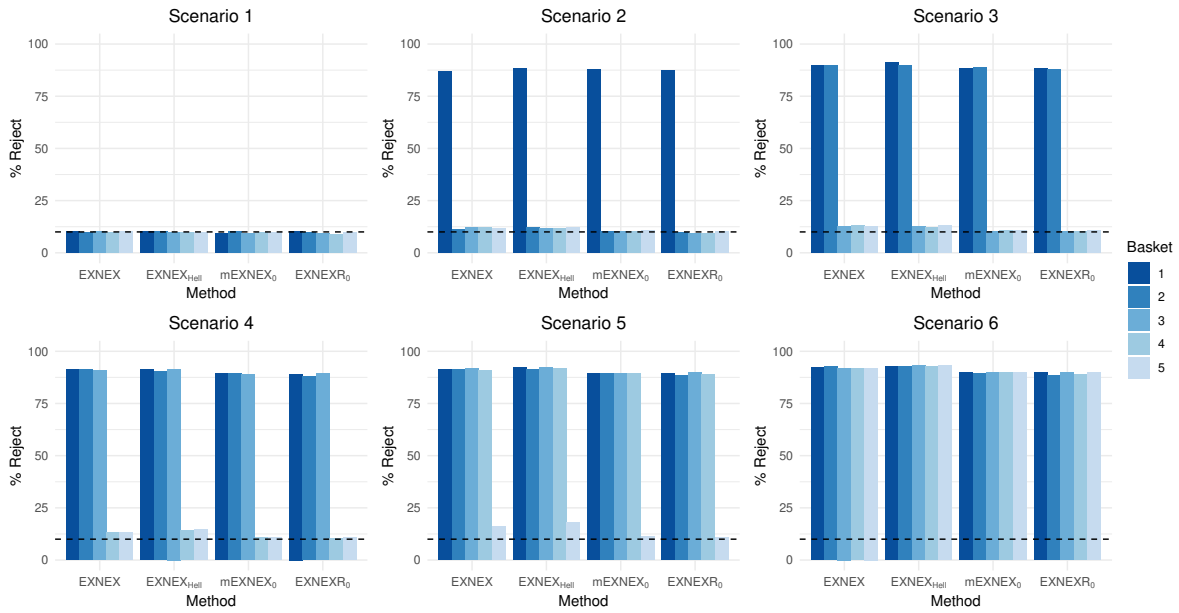


Figure A.5.1: The percentage of simulated data sets in which the null hypothesis was rejected in each basket under the four model settings outlined to compare the 1-step vs. 2-step mEXNEX$_c$ across the simulation settings provided in Table 2.3.1 in Chapter 2.

have far better error control at the cost of slightly lower statistical power. We would therefore not recommend the EXNEX$_{\text{Hell}}$ for use as it's performance is inferior to the other proposed modifications.

Only minimal differences in the mEXNEX$_c$ and EXNEXR$_c$ are observed here with the mEXNEX$_c$ method giving consistently greater power compared to the EXNEXR$_c$ method but only up to an increase of 0.5%. This increase occurs alongside insubstantial inflation in error rates. Although the difference in these two methods is only slight in this single study, we explore further trial settings to investigate the differences between the two proposals.

A.5.2 Varying the Study Design

When conducting simulation studies there are a few design parameters to consider, these include:

- The number of baskets, K, included in the study.

- The sample size, n_k, within the k baskets.
- The null and target response rate.

In the simulation study in Section A.5.1, the design parameters are specified to have $K = 5$ baskets with $n_k = 13$ patients in each of the $k = 1, \ldots, K$ baskets, while the null and target response rates are fixed at $q_0 = 0.15$ and $q_1 = 0.45$ respectively. These design parameters are in line with those of the VE-BASKET trial, which the simulation is motivated by. We now use this simulation as a reference setting, while we vary one of the three design parameters at a time to determine where differences in the mEXNEX$_c$ and EXNEXR$_c$, becomes more prominent. The different settings of design parameter alterations are provided in Table A.5.1.

Table A.5.1: Simulation settings for comparing the modified EXNEX models where we vary a single design parameter at a time.

Setting	No. Baskets (K)	Sample Size (n_k)	q_0	q_1
Reference	5	13	0.15	0.45
1	3	13	0.15	0.45
2	10	13	0.15	0.45
3	5	5	0.15	0.45
4	5	30	0.15	0.45
5	5	100	0.15	0.45
6	5	13	0.15	0.3
7	3	13	0.15	0.7

For the comparison of the design settings, the EXNEX$_{\text{Hell}}$ model was excluded due to it's inferior performance in the previous simulations. For both models that remove heterogeneous baskets we opt for two values of c, $c = 0.05$ and $c = 0.1$. The data scenario applied is one in which we have a ratio of two baskets with a response rate of q_1 to three baskets with a response rate of q_0.

Plotted in Figure A.5.2 are the rejection percentages for each basket and model under the 8 different simulation settings outlined in Table A.5.1. The first two of these settings vary the number of baskets from a very small number to moderately large,

whilst settings 3-5 cover different possible sample sizes per basket, a case where the number of patients is very small, a moderate number and then in setting 5 a larger number of patients. The final two settings vary the target response rate with one value being close to the null response rate and the second being fairly different from q_0.

Looking first at the reference model, an average of 0.7% increase in power is observed with just a 0.3% increase in type I error when using mEXNEX$_c$ over EXNEXR$_c$ when $c = 0.05$. However, as the number of baskets, K, increases, the difference in type I error rate is more evident. The increased inflation in error rates across all methods as K increases comes about as less baskets are being treated independent at the removal step. To treat a basket as independent we compute all pairwise difference in response rates, i.e. for basket i, we treat it as independent if $|X_i - X_j| > c$ for all $i \neq j$ where $X_i = Y_i/n_i$, with Y_i being the number of responses observed in basket i which consists of n_i patients. If we consider the case where the response rates are IID, then the probability a basket is treated as independent is $\mathbb{P}(|X_1 - X_2| > c)^{K-1}$ which is decreasing as K increases. As a result, when $K = 10$, borrowing will occur more frequently between heterogeneous baskets and hence the error rates inflate to anywhere between 12 and 18% dependent on method chosen (compared to 10.5 to 12.4% when $K = 5$). When $K = 3$, more baskets are treated as independent and thus differences in error rate and power are minimal.

More error rate inflation is observed when using mEXNEX$_c$ compared to EXNEXR$_c$ when $K = 10$ (up to a 2% increase when $c = 0.05$ and 3% when $c = 0.1$). This again comes down to fewer baskets being treated independently as despite this, the EXNEXR$_c$ model limits borrowing by fixing the prior probabilities at $\pi_k = 0.5$, whereas under the mEXNEX$_c$ model, these probabilities tend to be higher resulting in more borrowing from heterogeneous baskets and hence higher type I error rates.

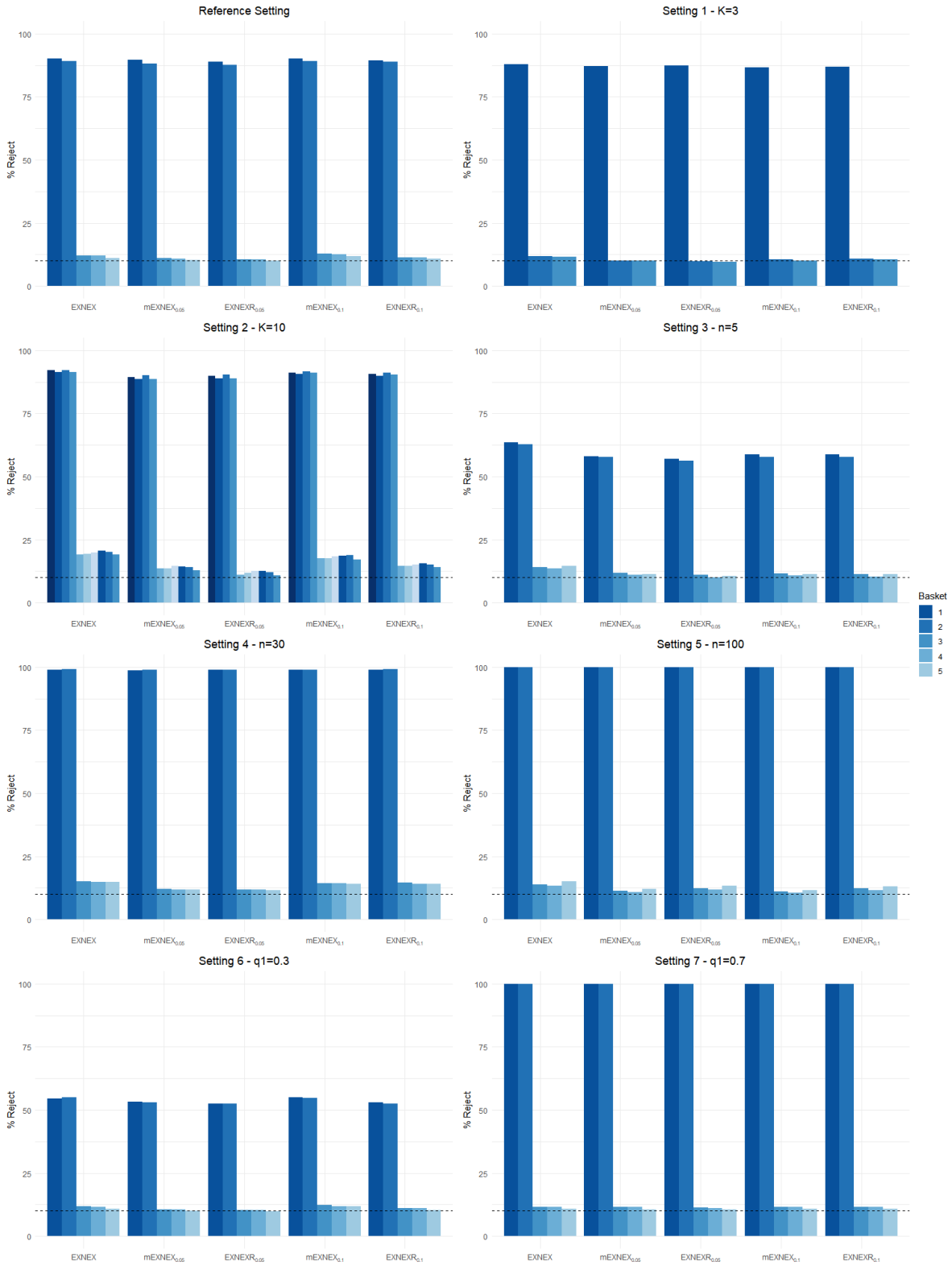


Figure A.5.2: The percentage of simulated data sets in which the null hypothesis was rejected in each basket under the eight trial design settings outlined in Table A.5.1, comparing the 1-step vs. 2-step mEXNEX$_c$ model.

Looking at the effect of sample size, we note that as the sample size increases, the Hellinger distance between two baskets decreases. These reduced Hellinger distances for both homogeneous and heterogeneous baskets, when averaged, result in smaller π_k values and hence less borrowing. In fact, as n_k increases to 100, these probabilities can actually fall below 0.5, so we therefore have a lower chance of borrowing between homogeneous baskets under the mEXNEX$_c$ model compared to the EXNEXR$_c$ model, and hence we observe better error control (11.45% compared to 12.48%). However, as the sample size increases, the increased certainty in estimates results in power tending towards 100% regardless of method and hence no improvement in power is observed. Even when n_k is small at just 5 patients in each basket, the difference in error rates is also relatively small at a 0.8% increase under the mEXNEX$_c$ model, whilst gaining on average 1.2% power when $c = 0.05$.

The final varied design parameter considered is the target response rate, q_1. When q_1 is closer to the null response rate, we observe minimal change in the type I error rate from the nominal 10% level, whereas, when $q_1 = 0.7$ this error rate rises to above 11% for both methods. Inflation in error rates is caused by a pull away from the true mean towards the common mean. When the target response rate is close to the null, this pull is less substantial compared to larger q_1 values, hence explaining the minimal difference in type I error rate when $q_1 = 0.3$. Similar error rates are observed under both methods, however the gain in power is greater under the mEXNEX$_c$ model at 53.2% compared to 52.5% under the EXNEXR$_{0.05}$ model when q_1 is closer to the null response rate. This is due to the Hellinger distance, and hence π_k values, being closer to 1 under the mEXNEX$_c$ model whereas under the EXNEXR$_c$ model these are fixed at 0.5.

From the above we conclude that, in general, the mEXNEX$_c$ model performs more favourably than EXNEXR$_c$ when the sample size is very small or very large and when the target response rate is closer to the null response rate. Whereas, the EXNEXR$_c$

model is preferred when the number of baskets increases and when the target response rate is very different to the null response rate. Overall, we would recommend the two-stage mEXNEX$_c$ method over the EXNEXR$_c$ model as a trial with fewer baskets and a target response rate closer to the null is more realistic. Although an argument could be made in some cases to use just the removal of heterogeneous baskets step.

A.6 Simulation Study for a Varied Number Baskets

In Section A.5.2 a simulation study was conducted with one parameter varied at a time, this section now focuses in on just one of those parameters: the number of baskets, denoted K. Two values of K are considered - $K = 3$ and $K = 10$ - with full simulation results for several data scenarios and under each of the borrowing methods provided in Sections A.6.1 and A.6.2.

For both cases a total of 10,000 simulation runs were used for each method and data scenario. The VE-BASKET trial remains as the motivating example, as such, the sample size in each basket is fixed at $n_k = 13$ for all $k = 1, \ldots, K$ with a null and target response rate of $q_0 = 0.15$ and $q_0 = 0.45$ respectively. Model specifications are consistent with those outlined in Appendix 2.6 of Chapter 2.

A.6.1 Simulation Study for $K = 3$ Baskets

Under $K = 3$ baskets, 8 data scenarios are considered and outlined in Table A.6.1. Scenarios 1-4 cover varying number of effective baskets from none to all 3, whilst scenarios 5-8 consist of cases where some baskets are marginally effective with a true response rate of $p_k = 0.35$.

The modified EXNEX model was calibrated as outlined in the procedure in Section 2.2.6 in Chapter 2 and results for two values, $c = 1/13$ and $c = 4/13$, are presented here. The CBHM was also tuned to get parameters $a = -1.390$ and $b = 3.674$. Calibrated

Table A.6.1: Simulation study scenarios for the $K = 3$ setting

	p_1	p_2	p_3
Scenario 1	0.15	0.15	0.15
Scenario 2	0.45	0.15	0.15
Scenario 3	0.45	0.45	0.15
Scenario 4	0.45	0.45	0.45
Scenario 5	0.35	0.15	0.15
Scenario 6	0.35	0.35	0.15
Scenario 7	0.45	0.35	0.15
Scenario 8	0.45	0.35	0.35

Δ_α values are provided in Table A.6.2. Simulation results are presented in table form in Table A.6.3, with rejection percentages also displayed in Figure A.6.1.

Table A.6.2: Calibrated Δ_α values for the $K = 3$ basket simulation.

	Δ_α
Independent	0.904
BHM	0.862
CBHM	0.895
BMA	0.873
EXNEX	0.894
mEXNEX$_{1/13}$	0.916
mEXNEX$_{4/13}$	0.891

Under data scenario 2, all methods give reasonably similar power values ranging from 86.9% to 88.4%. Both the BHM and BMA approach give the smallest power in this case with the mEXNEX$_{1/14}$ model presenting the highest. Error rates tend to be significantly smaller in the 3 basket case compared to the previous 5 basket simulation, as does the difference in performance between all methods. The BHM and BMA approach have inflated error rates at around 13%. Almost identical inflation is observed under the EXNEX and mEXNEX$_{4/14}$ model but the modified EXNEX model with $c = 1/13$ has lower error rates at 10.8% (compared to 11.5%). So the mEXNEX$_{1/13}$ model is appealing here for both it's error control and superior power value.

Across all scenarios 1-8 the standard EXNEX model and mEXNEX$_{4/14}$ model con-

tinue to perform almost identically with only marginal differences observed in terms of the type I error rate, power, FWER and percentage of all correct conclusions. However, when the cut-off is reduced to become more conservative at a value of $c = 1/13$, noticeable differences arise, including a slight reduction in error rates with a slight loss in power.

As in the previous simulation studies, the BHM and BMA procedure continue to inflate the type I error rate to an unacceptable level but in the 3 basket case, unlike the 5 basket case, tends to do little in terms of power improvement compared to the EXNEX models which possess superior error control. For example, under scenario 3 in which 2 of the 3 baskets are effective to treatment, power under the BHM is averaging at 90.9% with a type I error rate of 21.8%, whereas, under the EXNEX model power is 90.5% with a type I error rate of 11.6%. Therefore, one could argue that the minute power improvement can not justify the 10% increase in type I error rate. However, more improvement in power for the methods that borrow more strongly is observed in baskets with marginally effective response rates, most notably seen in scenario 8.

In a simulation akin to that in Section A.4, a further study was conducted within which the true response rate, p, was randomly generated within each simulation run. Results of this is presented in Figure A.6.2. Again, the standard EXNEX model and $\text{mEXNEX}_{4/13}$ model produce very similar results where both have type I error rate of 2.8%, but the $\text{mEXNEX}_{1/14}$ model has a slightly higher power at 85.9% compared to 85.7% under the EXNEX model. However, if the cut-off value c is reduced to 1/13 the error rate then becomes 2.7%, so a marginal decrease, with power 85.4%. Weighing up error control and power improvement one would favour the less conservative modified EXNEX approach or the standard EXNEX model as the error rates are all relatively similar but with power improvement over an independent analysis of about 2.5%.

All methods bar an independent analysis give approximately a 76% chance of making the correct conclusions across all baskets however, family-wise error rates vary.

Table A.6.3: Operating characteristics for a simulation consisting of $K = 3$ baskets.

	% Reject			% All Correct	FWER	% Reject			% All Correct	FWER
	Scenario 1					**Scenario 2**				
	0.15	**0.15**	**0.15**			**0.45**	**0.15**	**0.15**		
Independent	9.83	9.84	9.76	73.40	26.60	87.55	9.90	9.54	71.43	18.57
BHM	9.52	9.60	9.41	76.18	23.82	86.99	13.00	12.95	67.10	21.75
CBHM	9.79	9.69	9.52	76.05	23.95	87.30	12.82	12.67	66.77	21.95
BMA	9.37	9.71	9.39	76.74	23.26	86.86	14.51	14.12	66.22	22.48
EXNEX	9.97	10.22	9.84	73.93	26.07	88.03	11.56	11.49	68.34	21.66
$\text{mEXNEX}_{1/13}$	10.20	9.92	10.08	73.74	26.26	88.36	10.82	10.82	69.57	20.25
$\text{mEXNEX}_{4/13}$	9.94	10.14	9.84	74.04	25.96	88.31	11.52	11.42	68.62	21.55
	Scenario 3					**Scenario 4**				
	0.45	**0.45**	**0.15**			**0.45**	**0.45**	**0.45**		
Independent	87.78	87.44	9.77	69.27	9.77	87.94	87.21	87.84	67.97	
BHM	91.13	90.75	21.75	62.66	21.75	94.33	94.34	94.26	85.48	
CBHM	90.66	90.31	13.85	70.71	13.85	91.58	91.31	91.78	77.23	
BMA	91.53	91.27	21.98	62.72	21.98	94.47	94.23	94.59	85.70	
EXNEX	90.77	90.32	11.57	72.67	11.57	90.82	90.39	90.90	74.74	
$\text{mEXNEX}_{1/13}$	89.49	89.14	11.19	71.21	11.19	90.34	89.86	90.30	74.62	
$\text{mEXNEX}_{4/13}$	90.67	90.26	11.56	72.66	11.56	90.82	90.38	90.90	74.74	
	Scenario 5					**Scenario 6**				
	0.35	**0.15**	**0.15**			**0.35**	**0.35**	**0.15**		
Independent	67.9	9.88	9.76	55.48	18.64	67.83	66.33	9.71	40.52	9.71
BHM	66.34	12.60	12.52	50.19	21.19	73.04	71.63	19.94	38.04	19.94
CBHM	66.82	12.53	12.34	50.00	20.92	71.70	70.51	15.86	42.60	15.86
BMA	66.24	13.86	13.66	49.26	21.88	74.14	72.57	21.67	37.14	21.67
EXNEX	68.11	11.26	11.26	52.46	21.13	72.20	70.78	11.52	46.23	11.52
$\text{mEXNEX}_{1/13}$	68.80	10.69	10.41	53.66	19.73	70.06	68.63	11.06	45.09	11.06
$\text{mEXNEX}_{4/13}$	68.29	11.24	11.30	52.64	21.15	72.19	70.77	11.50	46.23	11.50
	Scenario 7					**Scenario 8**				
	0.45	**0.35**	**0.15**			**0.45**	**0.45**	**0.35**		
Independent	87.88	66.50	9.64	52.69	9.64	87.96	66.43	67.84	39.71	
BHM	90.64	72.50	20.68	48.48	20.68	93.59	79.47	80.26	65.73	
CBHM	90.12	71.68	14.67	54.89	14.67	91.92	74.29	75.54	54.26	
BMA	91.13	73.55	22.34	47.45	22.34	94.31	81.18	82.15	69.13	
EXNEX	90.40	71.26	11.53	57.47	11.53	90.71	71.46	72.85	47.69	
$\text{mEXNEX}_{1/13}$	89.41	69.69	11.01	56.12	11.01	89.81	69.86	71.30	47.64	
$\text{mEXNEX}_{4/13}$	90.30	71.31	11.55	57.47	11.55	90.67	71.44	72.82	47.59	

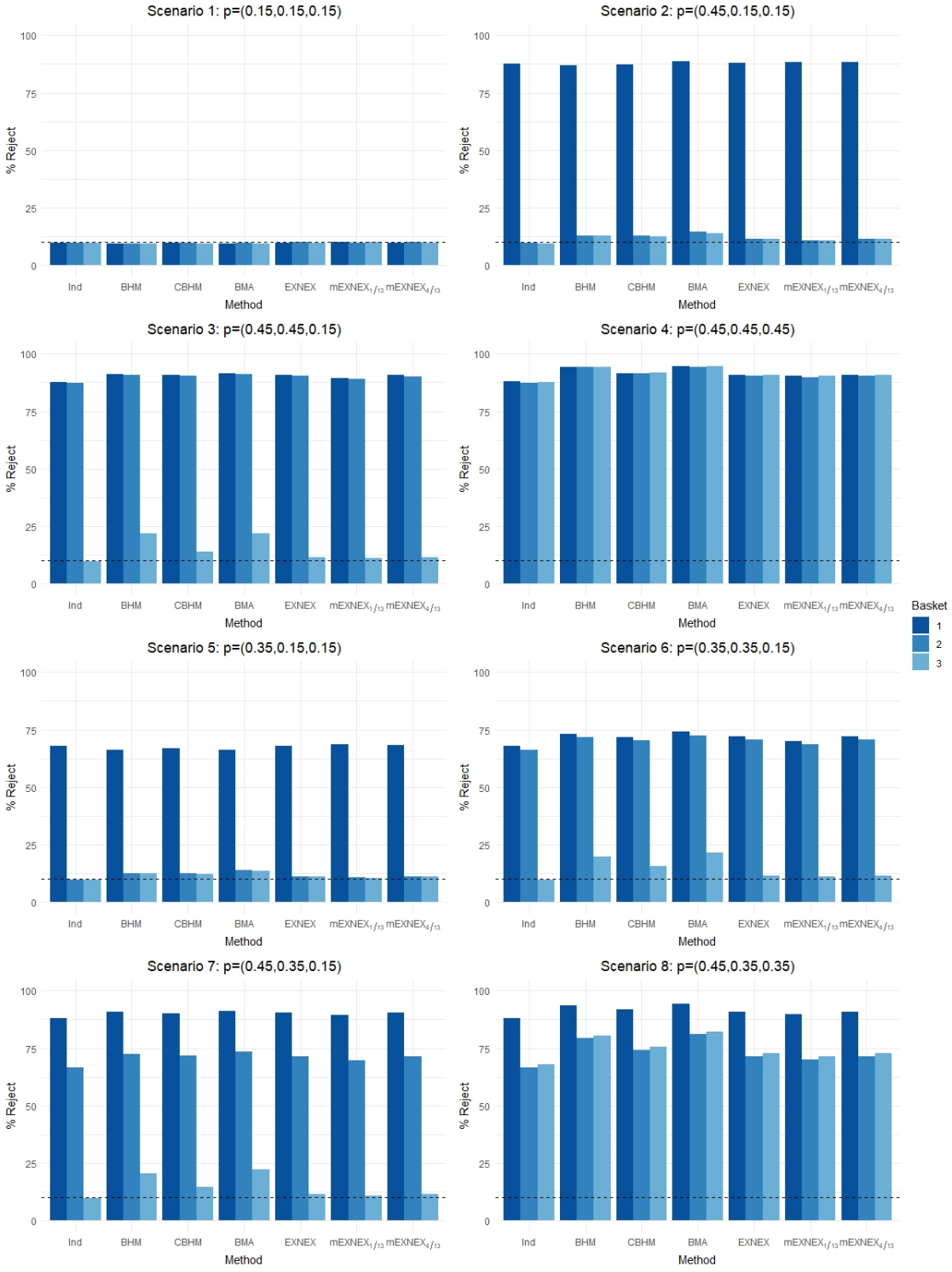


Figure A.6.1: Percentage of rejections of the null hypothesis for each information borrowing method under the $K = 3$ case.

The BHM and BMA approach give FWER values of around 6.2% whilst the standard EXNEX and mEXNEX$_{4/13}$ models have a FWER of 4.7% with the mEXNEX$_0$ model slightly lower at 4.6%.

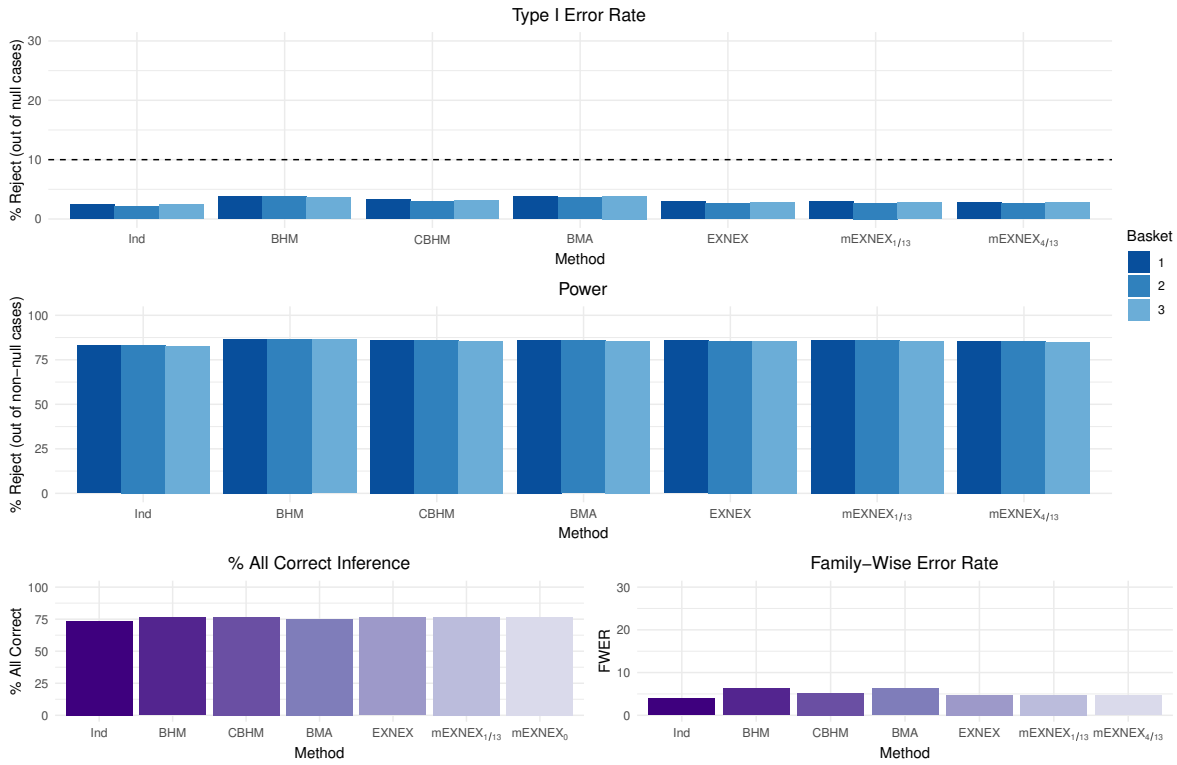


Figure A.6.2: Operating characteristics under varied truths for each information borrowing method for the $K = 3$ basket simulation study

To summarise, the effect of reducing the number of baskets is the slight reduction in power alongside a smaller inflation in error rates particularly for the BHM and BMA approach, however, the conclusions regarding method comparison appear to be very similar to that in the planned sample size case with $K = 5$ baskets.

A.6.2 Simulation Study for $K = 10$ Baskets

Under $K = 10$ baskets, 15 data scenarios are considered and outlined in Table A.6.4. Scenarios 1-11 cover varying number of effective baskets from 1-10, whilst scenarios 12-15 consist of cases where some baskets are marginally effective with a true response rate of $p_k = 0.35$.

Table A.6.4: Simulation study scenarios for the $K = 10$ setting.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_10
Scenario 1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 2	0.45	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 3	0.45	0.45	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 4	0.45	0.45	0.45	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 5	0.45	0.45	0.45	0.45	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 6	0.45	0.45	0.45	0.45	0.45	0.15	0.15	0.15	0.15	0.15
Scenario 7	0.45	0.45	0.45	0.45	0.45	0.45	0.15	0.15	0.15	0.15
Scenario 8	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.15	0.15	0.15
Scenario 9	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.15	0.15
Scenario 10	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.15
Scenario 11	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
Scenario 12	0.35	0.35	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
Scenario 13	0.35	0.35	0.35	0.35	0.35	0.15	0.15	0.15	0.15	0.15
Scenario 14	0.45	0.45	0.45	0.35	0.35	0.35	0.15	0.15	0.15	0.15
Scenario 15	0.45	0.45	0.35	0.35	0.35	0.35	0.35	0.35	0.15	0.15

The modified EXNEX model was calibrated as outlined in the procedure in Section 2.2.6 in Chapter 2 and results of two calibrated cut-off values are presented here - $c = 0$ and $c = 1/13$. The CBHM was also tuned to get parameters $a = -23.475$ and $b = 10.963$. Calibrated Δ_α values are provided in Table A.6.5.

Full results are presented in Tables A.6.6, A.6.7 and A.6.8, with the hypothesis rejection percentages also displayed in Figures A.6.3 and A.6.4.

Consider scenario 2 where a single basket is heterogeneous and effective, the results are similar to that in the $K = 5$ basket case. Again, substantial inflation in the type I error rate is observed under the BHM and BMA approach, with an averaged error rate of 17.2% and 12.7% respectively. Under the $K = 5$ basket case, a similar scenario with

Table A.6.5: Calibrated Δ_α values for the $K = 10$ basket simulation.

	Δ_α
Independent	0.904
BHM	0.797
CBHM	0.896
BMA	0.844
EXNEX	0.833
mEXNEX$_0$	0.868
mEXNEX$_{1/13}$	0.841

a single heterogeneous basket resulted in an error rate of 16.9% and 13.2% respectively - these values are very similar indicating that increasing the number of baskets does little to eliminate error rate inflation under the two models that demonstrate the worst performance, particularly as in these scenarios where the power is substantially lower than an independent analysis (82.2% under the BHM compared to 88.3%).

Looking at other methods under the same scenario, the CBHM also has inflated error rates at approximately 12%. This contradicts the calibration nature of this model that takes a 'strong' definition of heterogeneity in that: if a single or multiple baskets have a heterogeneous response, then all are deemed heterogeneous so analysed as independent. In this scenario, there is clear heterogeneity thus it would be expected that the CBHM performs similarly to the independent model. This is not the case, leading to the conclusion that perhaps the calibration was slightly off.

The EXNEX model under scenario 2 also has fairly substantial error rates at 11.9% with a power of 87%, whereas, under the more conservative modified EXNEX approach with $c = 0$, error rates are around 10.5%, so close to the nominal level, with a power of 87.5% - an improvement over the standard EXNEX model. If the cut-off was increased to $c = 1/13$, error rates increase to 11.3% with 87.4% power. Bringing together these results, under scenario 2, of the borrowing models the mEXNEX$_0$ model has both the highest power and best error control.

Moving on to cases in which multiple baskets are effective to the treatment (as in

scenarios 3-10), the pattern of results described above hold in that the BHM and BMA inflate error rates, the mEXNEX$_c$ models have better error control than the standard EXNEX model, particularly when c is more conservative, whilst still improving power over an independent analysis. These are also the same conclusions drawn from the 5 basket simulation study presented in Chapter 2.

One can see that the error rate gets far more substantial across all methods as the ratio of effective to ineffective baskets increases with the inflation greater than that in the 5 basket case. For example, under $K = 5$ the maximum error rate for the BHM is 42.1%, whereas in the $K = 10$ case this is 76.8% which occurs under scenarios 5 and 10 respectively. In scenario 10 for $K = 10$, just one basket is ineffective with a true response rate of 0.15 whilst the other 9 are effective with a higher response at 0.45. The presence of 9 effective basket compared to just 4 in the $K = 5$ case causes a larger pull up towards the common mean for the single heterogeneous basket, hence the greater inflation. This holds for all methods, with maximum error rates uniformly increasing from the $K = 5$ to $K = 10$ case. However, although this error inflation is observed, power is improved across all methods, this is unsurprising due to the additional certainty we gain from the extra 5 baskets included in the study.

Under the four scenarios in which a number of baskets have a marginally effective response rate of 35%, all borrowing methods show a substantial gain in power compared to an independent analysis, particularly for the marginally effective baskets. But inflation in error rates continues to be an issue.

As in Section A.4, a further simulation was conducted within which the truth vector, p, is varied within each simulation. The results are provided in Figure A.6.5. Unlike in the $K = 3$ basket case, more substantial differences are observed in the standard EXNEX model and the modified EXNEX approaches. The standard EXNEX model has error rates and power of 4.6% and 88.3% respectively, the mEXNEX$_0$ model has an average error rate of 3.3% and power 86.2% whilst the mEXNEX$_{1/13}$ model has 4.4%

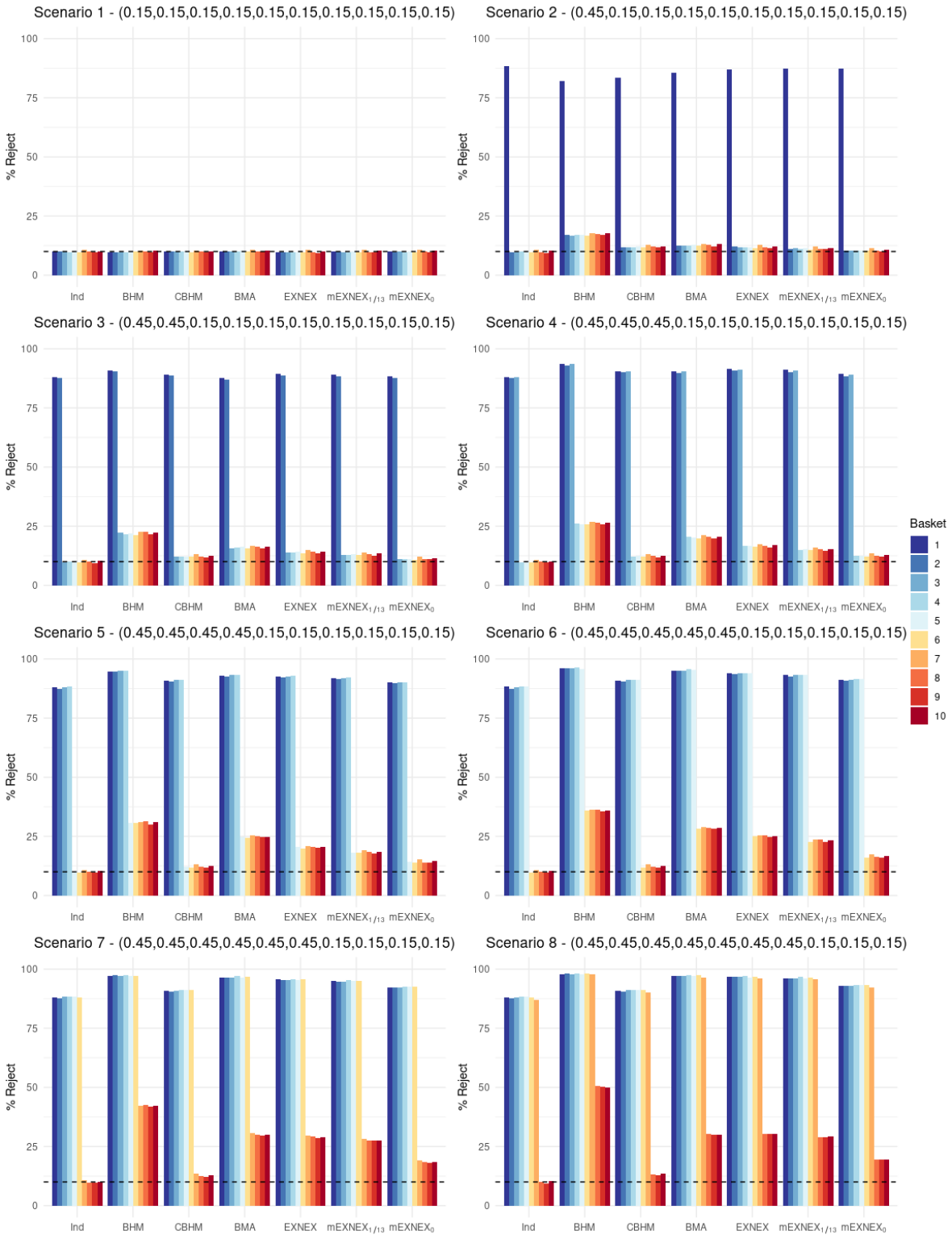


Figure A.6.3: Percentage of rejections of the null hypothesis for each information borrowing method under the $K = 10$ case across scenarios 1-8.

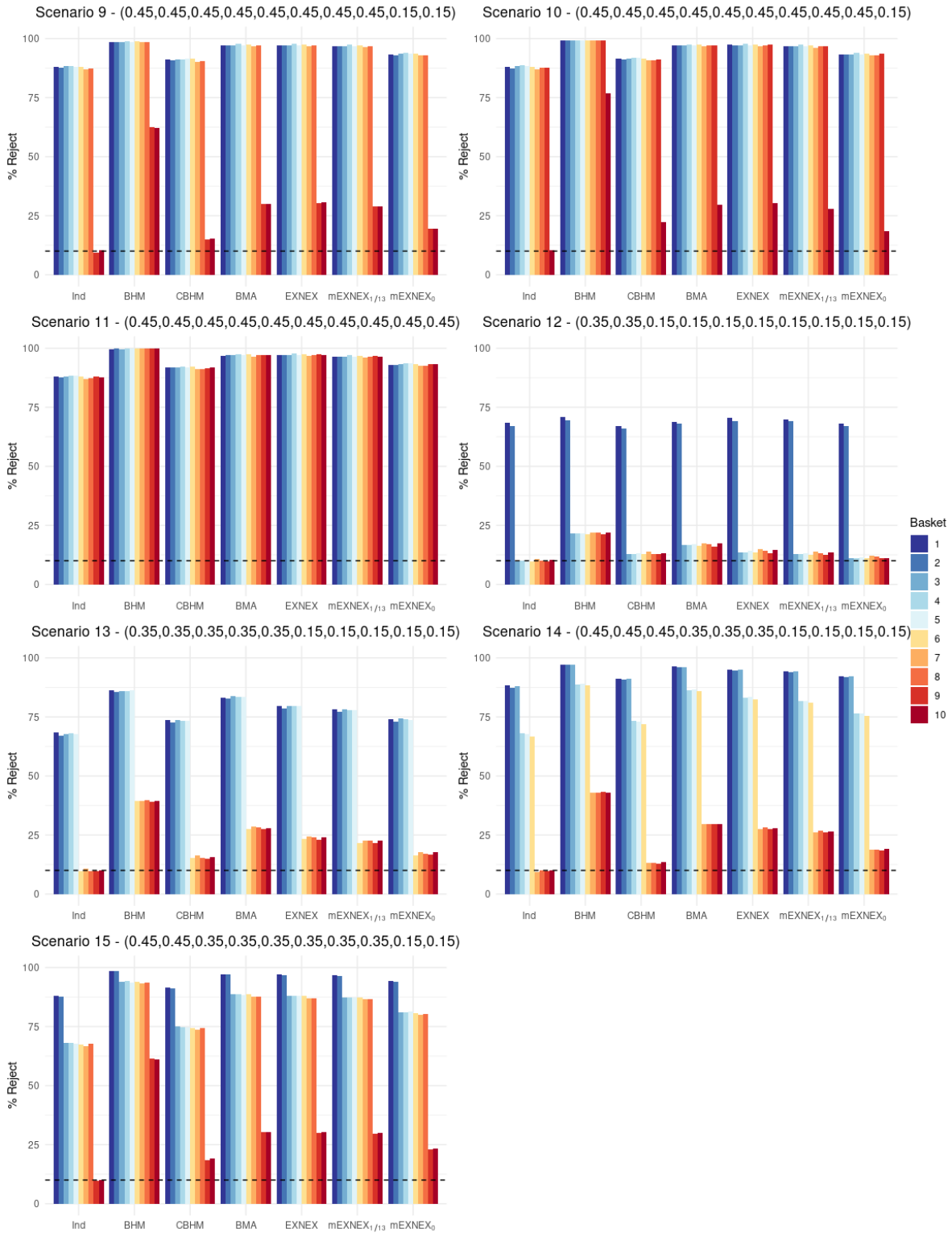


Figure A.6.4: Percentage of rejections of the null hypothesis for each information borrowing method under the $K = 10$ case across scenarios 9-15.

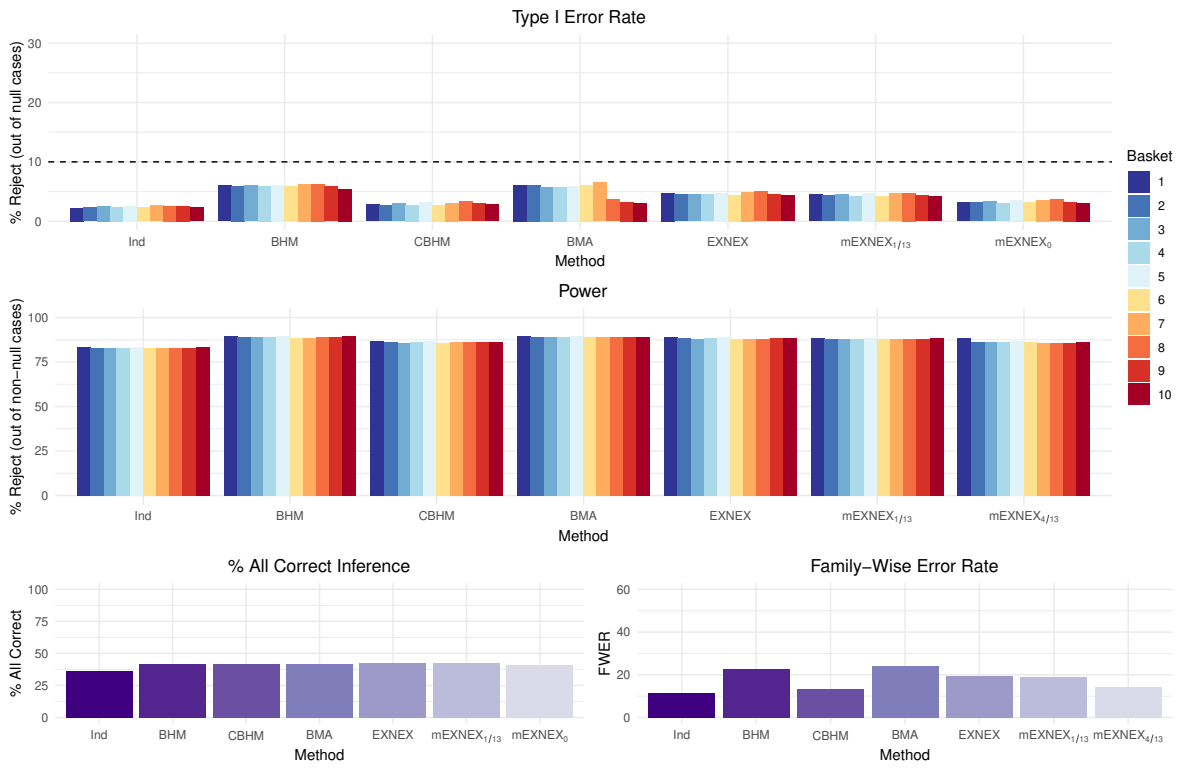


Figure A.6.5: Operating characteristics for comparison of information borrowing methods under varied truths for the $K = 10$ basket simulation study

error rate and 88.1% power. All have power improvement over an independent analysis which has an average power of 83%, thus even under the most conservative modified EXNEX approach, power is improved by around 6.2%.

Then looking at the percentage of times correct inference was made across all 10 baskets, highest values were observed under the EXNEX model, mEXNEX$_{1/13}$ model, a BMA approach and the BHM at around 42%. The mEXNEX$_0$ model has value of 41.1% for all correct inference but the mEXNEX$_0$ model had 5% lower family-wise error rate compared to the standard EXNEX model. Weighing up both FWER and all correct inference, the mEXNEX$_0$ model appears optimal with FWER closer to that of the independent analysis and CBHM, whilst giving 3% higher percentage of correct inference compared to an independent analysis.

To summarise, this study confirms the conclusions drawn from the $K = 5$ basket simulation study presented in Chapter 2 whilst also highlighting that the larger number

of baskets, although improving certainty of estimates and power, causes an even less favourable type I error rates.

Table A.6.6: Operating characteristics for a simulation based on $K = 10$ baskets with a sample size of $n_k = 13$ in each (scenarios 1-6)

	% Reject										% All Correct	FWER
Sample Size	13	13	13	13	13	13	13	13	13	13		
Scenario 1	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	9.89	9.96	9.89	9.83	10.06	9.81	10.74	10.03	9.84	10.18	35.07	0.649
BHM	9.73	9.87	9.71	9.81	10.19	9.98	10.39	9.89	9.80	10.31	65.32	0.347
CBHM	10.21	10.14	9.88	9.91	10.19	9.80	10.55	10.14	9.75	10.09	59.35	0.407
BMA	10.08	10.17	9.99	10.14	10.39	10.16	10.85	9.91	10.06	10.51	54.78	0.452
EXNEX	9.67	9.59	9.71	9.74	9.78	9.60	10.62	9.53	9.50	10.18	49.15	0.509
$\text{mEXNEX}_{1/13}$	9.94	9.92	9.81	9.82	9.95	9.73	10.79	9.59	9.76	10.40	44.86	0.551
mEXNEX_0	10.05	9.81	9.95	9.81	9.97	9.60	10.82	9.96	9.87	10.25	36.39	0.636
Scenario 2	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.34	9.80	9.84	9.88	9.97	9.65	10.71	9.85	9.51	10.40	34.97	0.608
BHM	82.15	16.95	16.85	17.12	17.17	16.77	17.67	17.45	16.90	17.78	25.42	0.596
CBHM	83.48	11.79	11.90	11.93	12.01	11.68	12.71	12.01	11.69	12.26	28.57	0.589
BMA	85.53	12.56	12.61	12.49	12.98	12.32	13.36	12.78	12.19	13.08	42.26	0.478
EXNEX	86.99	12.00	11.66	11.77	11.95	11.52	12.78	11.94	11.35	12.31	34.97	0.572
$\text{mEXNEX}_{1/13}$	87.43	11.26	11.29	11.17	11.35	10.81	12.09	11.06	10.96	11.50	35.93	0.570
mEXNEX_0	87.45	10.33	10.50	10.41	10.49	10.32	11.37	10.40	10.09	10.70	34.27	0.598
Scenario 3	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.05	87.74	9.77	10.01	10.05	9.72	10.64	10.13	9.39	10.34	33.75	0.564
BHM	90.73	90.48	22.14	21.75	21.76	21.30	22.63	22.55	21.67	22.37	24.30	0.670
CBHM	89.05	88.74	12.24	12.00	12.23	12.01	13.15	12.26	11.84	12.58	29.69	0.606
BMA	87.82	86.99	15.79	16.05	16.28	15.75	16.84	16.34	15.51	16.49	31.44	0.528
EXNEX	89.52	88.72	13.87	13.80	14.18	13.59	14.99	14.28	13.51	14.28	28.73	0.605
$\text{mEXNEX}_{1/13}$	89.12	88.22	13.01	12.83	13.21	12.77	13.75	13.14	12.50	13.39	29.01	0.595
mEXNEX_0	88.35	87.56	11.26	11.07	11.24	10.91	11.97	11.22	11.06	11.43	32.04	0.563
Scenario 4	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.05	87.51	87.87	9.69	9.88	9.67	10.77	9.91	9.63	10.01	32.76	0.518
BHM	93.52	93.02	93.48	26.27	25.73	25.61	26.74	26.51	25.85	26.35	21.91	0.696
CBHM	90.47	90.05	90.61	12.21	12.33	12.09	13.22	12.37	11.96	12.56	31.38	0.574
BMA	90.39	89.72	90.44	20.48	20.12	19.95	21.08	20.61	19.71	20.67	24.62	0.585
EXNEX	91.43	90.79	91.30	16.55	16.56	16.26	17.45	16.86	16.02	16.90	20.29	0.593
$\text{mEXNEX}_{1/13}$	90.99	90.09	90.75	14.83	15.13	14.84	15.92	15.33	14.63	15.33	29.84	0.580
mEXNEX_0	89.25	88.46	89.10	12.33	12.59	12.20	13.46	12.43	12.09	12.88	30.93	0.543
Scenario 5	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.16	87.39	88.09	88.24	10.06	9.64	10.81	10.04	9.75	10.22	31.38	0.472
BHM	94.76	94.59	94.89	95.13	30.73	30.71	31.07	31.27	30.03	31.01	17.54	0.733
CBHM	90.79	90.37	90.99	91.13	12.18	11.83	13.11	12.11	11.81	12.36	31.92	0.524
BMA	92.98	92.45	93.12	93.10	24.87	24.51	25.38	25.02	24.58	24.81	19.50	0.655
EXNEX	92.50	92.23	92.51	92.80	20.40	19.92	20.99	20.53	20.02	20.43	27.11	0.587
$\text{mEXNEX}_{1/13}$	91.90	91.45	91.83	92.08	18.18	17.99	19.14	18.30	17.86	18.39	29.14	0.557
mEXNEX_0	90.16	89.77	90.19	90.22	14.41	13.90	15.27	14.02	13.72	14.56	29.98	0.520
Scenario 6	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.44	87.38	88.15	88.38	88.49	9.61	10.79	10.05	9.63	10.37	31.21	0.413
BHM	96.04	96.06	96.16	96.34	95.88	35.96	36.30	36.24	35.57	35.98	15.50	0.766
CBHM	90.88	90.50	91.04	91.17	91.16	11.86	13.09	12.08	11.74	12.45	32.82	0.467
BMA	95.17	94.97	95.15	95.55	95.28	28.31	28.96	28.60	28.07	28.45	14.57	0.742
EXNEX	93.96	93.49	93.90	94.02	93.92	24.91	25.59	25.46	24.72	25.09	20.22	0.641
$\text{mEXNEX}_{1/13}$	93.17	92.73	93.14	93.32	93.34	22.62	23.55	23.67	22.77	23.18	23.05	0.594
mEXNEX_0	91.31	90.80	91.18	91.54	91.39	16.01	17.29	16.34	15.90	16.68	29.14	0.504

Table A.6.7: Operating characteristics for a simulation based on $K = 10$ baskets with a sample size of $n_k = 13$ in each (scenarios 7-12)

	% Reject										% All Correct	FWER
Sample Size	13	13	13	13	13	13	13	13	13	13		
Scenario 7	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.13	87.51	88.33	88.20	88.24	88.05	10.64	9.74	9.55	10.05	30.97	0.344
BHM	97.07	97.32	97.18	97.47	97.07	97.24	42.30	42.55	41.80	42.14	17.43	0.767
CBHM	90.89	90.51	90.98	91.23	91.09	91.06	13.47	12.55	12.14	12.80	33.58	0.399
BMA	96.49	96.42	96.55	96.99	96.45	96.67	30.74	30.14	29.75	29.86	18.46	0.746
EXNEX	95.59	95.23	95.33	95.82	95.38	95.67	29.66	29.26	28.63	28.88	16.83	0.709
$\text{mEXNEX}_{1/13}$	94.84	94.69	94.61	95.19	94.85	95.02	28.09	27.61	27.38	27.64	17.91	0.667
mEXNEX_0	92.19	92.06	92.15	92.59	92.48	92.43	19.17	18.41	18.00	18.46	29.39	0.478
Scenario 8	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**	**0.15**		
Independent	88.06	87.51	87.95	88.20	88.48	87.99	87.06	9.92	9.52	10.28	29.76	0.268
BHM	97.80	98.00	97.94	98.19	97.78	98.00	97.65	50.60	50.32	50.04	20.19	0.737
CBHM	90.98	90.57	91.07	91.24	91.19	91.07	89.99	13.09	12.81	13.44	34.78	0.313
BMA	97.07	96.98	97.01	97.51	96.94	97.37	96.56	30.38	30.07	30.00	27.89	0.653
EXNEX	96.70	96.62	96.70	97.18	96.51	96.92	96.22	30.40	30.16	30.25	26.54	0.654
$\text{mEXNEX}_{1/13}$	96.11	96.15	96.19	96.71	96.11	96.51	95.82	29.08	29.01	29.12	27.14	0.625
mEXNEX_0	93.03	92.81	92.92	93.31	93.37	93.14	92.21	19.41	19.42	19.48	32.31	0.421
Scenario 9	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**	**0.15**		
Independent	87.94	87.53	88.31	88.41	88.10	88.05	87.00	87.45	9.46	10.32	29.22	0.187
BHM	98.51	98.53	98.51	98.81	98.45	98.73	98.46	98.49	62.33	62.06	17.69	0.760
CBHM	91.08	90.69	91.28	91.32	91.34	91.34	90.24	90.46	15.04	15.44	34.04	0.242
BMA	97.13	97.11	97.16	97.69	96.99	97.51	96.69	97.19	29.99	29.95	39.12	0.506
EXNEX	97.20	97.11	97.21	97.70	97.08	97.50	96.78	97.21	30.40	30.50	38.78	0.515
$\text{mEXNEX}_{1/13}$	96.70	96.67	96.83	97.38	96.70	97.14	96.30	96.79	28.86	28.90	39.34	0.487
mEXNEX_0	93.20	92.98	93.48	93.86	93.47	93.50	92.88	93.00	19.38	19.45	36.18	0.327
Scenario 10	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.15**		
Independent	88.19	87.37	88.26	88.59	88.25	88.15	87.13	87.59	87.74	10.23	28.38	0.102
BHM	99.18	99.24	99.31	99.28	99.26	99.28	99.20	99.21	99.34	76.82	18.83	0.768
CBHM	91.53	91.21	91.61	91.88	91.70	91.68	90.71	90.91	91.30	22.24	29.94	0.222
BMA	97.06	97.07	97.09	97.62	97.02	97.46	96.72	97.14	97.28	29.62	54.14	0.296
EXNEX	97.29	97.22	97.24	97.77	97.12	97.55	96.85	97.26	97.42	30.49	54.31	0.305
$\text{mEXNEX}_{1/13}$	96.65	96.66	96.75	97.33	96.67	97.10	96.21	96.78	96.92	28.02	54.14	0.280
mEXNEX_0	93.2	93.11	93.35	93.97	93.40	93.50	92.82	92.95	93.64	18.45	41.18	0.185
Scenario 11	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**		
Independent	88.03	87.60	88.06	88.52	88.24	88.06	87.08	87.36	88.00	87.82	27.50	
BHM	99.64	99.79	99.72	99.78	99.74	99.79	99.83	99.78	99.76	99.82	98.03	
CBHM	91.72	91.70	91.84	92.38	91.84	92.12	91.19	91.27	91.67	91.95	57.50	
BMA	96.88	97.00	96.97	97.50	96.95	97.41	96.56	97.05	97.22	97.03	74.24	
EXNEX	97.26	97.21	97.27	97.77	97.13	97.56	96.85	97.28	97.43	97.20	76.07	
$\text{mEXNEX}_{1/13}$	96.27	96.35	96.45	96.96	96.35	96.83	95.97	96.43	96.70	96.40	68.66	
mEXNEX_0	92.96	92.85	93.19	93.58	93.49	93.36	92.57	92.69	93.41	93.11	46.12	
Scenario 12	**0.35**	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	68.34	66.92	9.82	9.89	9.87	9.79	10.72	9.94	9.69	10.25	20.25	0.567
BHM	70.88	69.50	21.52	21.58	21.51	21.13	21.84	21.79	21.11	21.91	13.30	0.609
CBHM	66.90	66.10	12.89	12.98	13.02	12.88	13.75	12.94	12.88	13.35	17.28	0.549
BMA	68.66	67.95	16.58	16.60	16.90	16.30	17.43	17.10	16.10	17.33	14.68	0.544
EXNEX	70.49	69.26	13.69	13.66	14.21	13.62	14.87	14.11	13.29	14.69	16.91	0.594
$\text{mEXNEX}_{1/13}$	69.93	69.02	12.96	12.77	13.08	12.65	13.91	13.16	12.46	13.44	17.33	0.581
mEXNEX_0	68.10	67.00	21.02	11.23	11.27	11.11	12.04	11.63	11.02	11.24	19.29	0.552

Table A.6.8: Operating characteristics for a simulation based on $K = 10$ baskets with a sample size of $n_k = 13$ in each (scenarios 13-15)

	% Reject										% All Correct	FWER
Sample Size	13	13	13	13	13	13	13	13	13	13		
Scenario 13	**0.35**	**0.35**	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	68.46	67.02	67.76	68.16	67.79	9.53	10.74	9.87	9.62	10.05	8.82	0.410
BHM	86.13	85.42	85.97	86.07	86.36	39.53	39.31	39.61	39.25	39.33	6.42	0.750
CBHM	73.76	72.59	73.82	73.47	73.35	15.35	16.28	15.44	15.01	15.59	10.58	0.475
BMA	83.26	82.73	83.70	83.44	83.48	27.63	28.44	28.22	27.52	28.01	5.49	0.730
EXNEX	79.66	78.58	79.76	79.68	79.64	23.43	24.52	24.10	23.13	24.20	5.84	0.618
mEXNEX$_{1/13}$	78.38	77.28	78.20	78.05	77.95	21.60	22.71	22.46	21.67	22.68	6.38	0.581
mEXNEX$_0$	74.20	73.13	74.30	74.05	73.64	16.46	17.60	16.97	16.64	17.60	8.89	0.501
Scenario 14	**0.45**	**0.45**	**0.45**	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**	**0.15**	**0.15**		
Independent	88.42	87.47	88.13	68.06	67.87	66.54	9.44	10.03	9.82	10.06	13.82	0.339
BHM	96.94	96.99	97.14	88.71	89.20	88.28	43.01	43.03	43.09	42.93	11.64	0.754
CBHM	91.13	90.79	91.26	73.36	73.12	71.76	13.28	13.33	12.91	13.61	16.89	0.396
BMA	96.39	96.21	96.22	86.40	86.58	85.83	29.49	29.79	29.58	29.73	13.27	0.731
EXNEX	95.06	94.74	94.94	83.16	83.48	82.55	27.68	28.11	27.59	27.96	10.54	0.675
mEXNEX$_{1/13}$	94.46	94.02	94.33	81.74	81.84	81.08	26.24	26.78	26.26	26.58	9.54	0.636
mEXNEX$_0$	92.07	91.91	92.07	76.64	76.35	75.30	18.62	18.69	18.44	19.02	14.92	0.476
Scenario 15	**0.45**	**0.45**	**0.35**	**0.35**	**0.35**	**0.35**	**0.35**	**0.35**	**0.15**	**0.15**		
Independent	88.02	87.77	68.18	67.92	67.62	67.49	66.63	67.74	9.54	10.20	6.58	0.187
BHM	98.58	98.62	93.90	94.20	93.70	93.88	93.34	93.64	61.48	61.00	9.95	0.746
CBHM	91.63	91.24	75.06	74.19	74.75	74.31	73.55	74.24	18.52	19.02	8.54	0.273
BMA	97.22	97.12	88.65	88.71	88.50	88.63	87.68	87.82	30.42	30.47	22.09	0.515
EXNEX	96.94	96.73	87.99	87.99	87.93	87.95	87.09	87.04	30.14	30.21	22.03	0.509
mEXNEX$_{1/13}$	96.59	96.54	87.37	87.31	87.23	87.25	86.51	86.53	29.27	29.82	21.89	0.502
mEXNEX$_0$	94.29	93.93	80.98	81.09	81.28	80.66	80.02	80.43	22.97	23.40	14.73	0.388

Appendix B

Supporting Information: How to Add Baskets to an Ongoing Basket Trial with Information Borrowing

B.1 Fixed Scenario Simulation Results Under the RCaP

A further 10 scenarios were considered in the fixed scenario simulation study presented in Chapter 3. All 16 scenarios considered are presented in Table B.1.1, where efficacy criteria were calibrated using RCaP. Tables B.1.2-B.1.5 present full simulation results with all operating characteristics. Figure B.1.1 plots the relative difference between calibration approaches under the same scenarios and Figure B.1.2 plots the results of the additional 10 data scenarios.

Comparing calibration approaches, scenarios 7 and 8 have similar findings to scenarios 1 and 2 presented in Chapter 3, in terms of inflated error rates when efficacy criteria are calibrated under the global null. However, error rates under the RCaP are also inflated but to a much lesser extent (e.g. 42.4% compared to 84.7% in existing baskets

Table B.1.1: Full list of 16 simulation study scenarios: Vectors of response rates used within the simulation study to compare approaches for adding a basket..

	p_1	p_2	p_3	p_4	p_5		p_1	p_2	p_3	p_4	p_5
Scenario 1	0.2	0.2	0.2	0.2	0.2	Scenario 9	0.4	0.4	0.2	0.2	0.4
Scenario 2	0.4	0.2	0.2	0.2	0.2	Scenario 10	0.4	0.4	0.4	0.2	0.4
Scenario 3	0.4	0.4	0.2	0.2	0.2	Scenario 11	0.3	0.2	0.2	0.2	0.2
Scenario 4	0.4	0.4	0.4	0.2	0.2	Scenario 12	0.3	0.3	0.2	0.2	0.2
Scenario 5	0.4	0.4	0.4	0.4	0.2	Scenario 13	0.3	0.2	0.2	0.2	0.3
Scenario 6	0.4	0.4	0.4	0.4	0.4	Scenario 14	0.3	0.3	0.2	0.2	0.3
Scenario 7	0.2	0.2	0.2	0.2	0.4	Scenario 15	0.4	0.3	0.2	0.2	0.3
Scenario 8	0.4	0.2	0.2	0.2	0.4	Scenario 16	0.4	0.3	0.3	0.2	0.3

under PL2(a)) compared to the calibration under the global null. Under scenario 9, error rates are increased by up to 81.4% of the nominal 10% level compared to a 20.9% increase under the RCaP.

Scenarios 11, 12, 13 and 14 are parallel to the results under scenarios 2, 7, 9 and 10 respectively. The main differences lying in the results of power, as the baskets now have a marginally effective response rate, making it more difficult to distinguish between an effective and ineffective treatment effect. In all cases, power under the RCaP is lower due to the more conservative Δ_k cut-off values, however, this came with a reduced error rate across all 4 of these scenarios and 4 approaches. For instance, under scenario 14 error rates under the calibration under the global null have a relative increase of 54% over the nominal 10% level, compared to error rates controlled at or below 10% under the RCaP. Similar results are found under scenarios 14-16 in which baskets are a combination of effective, marginally effective and ineffective.

Comparing approaches for addition of a basket, under scenario 3, error in the new basket under IND is the lowest, but the maximum inflation of the type I error rate over the 10% nominal level is only 1.1% (under PL1(a) and PL2(a), which are equivalent when analysing new baskets). Scenario 4 shows consistent power in all non-null existing baskets across all 4 approaches, all above the targeted 80% level. The UNPL approach demonstrates marginally lower power than other methods. Basket 4 has type I error

rate which is slightly higher under IND and PL2(a). This is due to the common mean across baskets 1-4 being higher than across all baskets, due to the new basket being ineffective to treatment. Hence, fewer false rejections should be made under UNPL and PL1(a).

When the new basket is effective, under scenarios 9 and 10, substantial improvements in power are observed in the new basket when information borrowing is utilised. In scenarios 9 and 10, as a number of existing baskets are also effective, borrowing information between all baskets substantially improves power in the new. Under the IND and PL1(a) approaches, error rates in the existing baskets are slightly higher at around 12.5% in scenario 9 and 16.5% in scenario 10. Whereas, error rates in IND and PL2(a) have an error rate of 14.3% at the cost of reduced power in other baskets.

Results of scenarios 11, 12 and 13 correlate to those of scenarios 2, 7 and 6 respectively but with marginally effective rather than effective true response rates. This gives lower power across baskets but similar patterns in results. Scenario 14 does differ from the results of scenario 9, with IND now producing the highest power in the new basket.

Scenarios 15 and 16 have a combination of effective, marginally effective and ineffective baskets. For existing baskets, those that are marginally effective have similar power values across all approaches under scenario 15 of around 42.7%, however, more variation is observed under scenario 16, with PL1(a) producing a power of 47%, which is higher than UNPL with power 46.1% and IND and PL2(a) at 44.7%. But for the single effective basket all approaches give similar power values ranging from 78.7%-79.0% under scenario 15 and 80.9%-81.8% under scenario 16. Error rates in the existing baskets are higher under the IND and PL1(a) approaches for both scenarios as the posterior probabilities are pulled up via borrowing from the effective new baskets. The maximum error rate across both scenarios is 13.5%. For the new basket UNPL, PL1 and PL2 are almost identical in power, with the IND approach giving lower power under scenario 16 due to the lack of borrowing from the mostly homogeneous existing baskets.

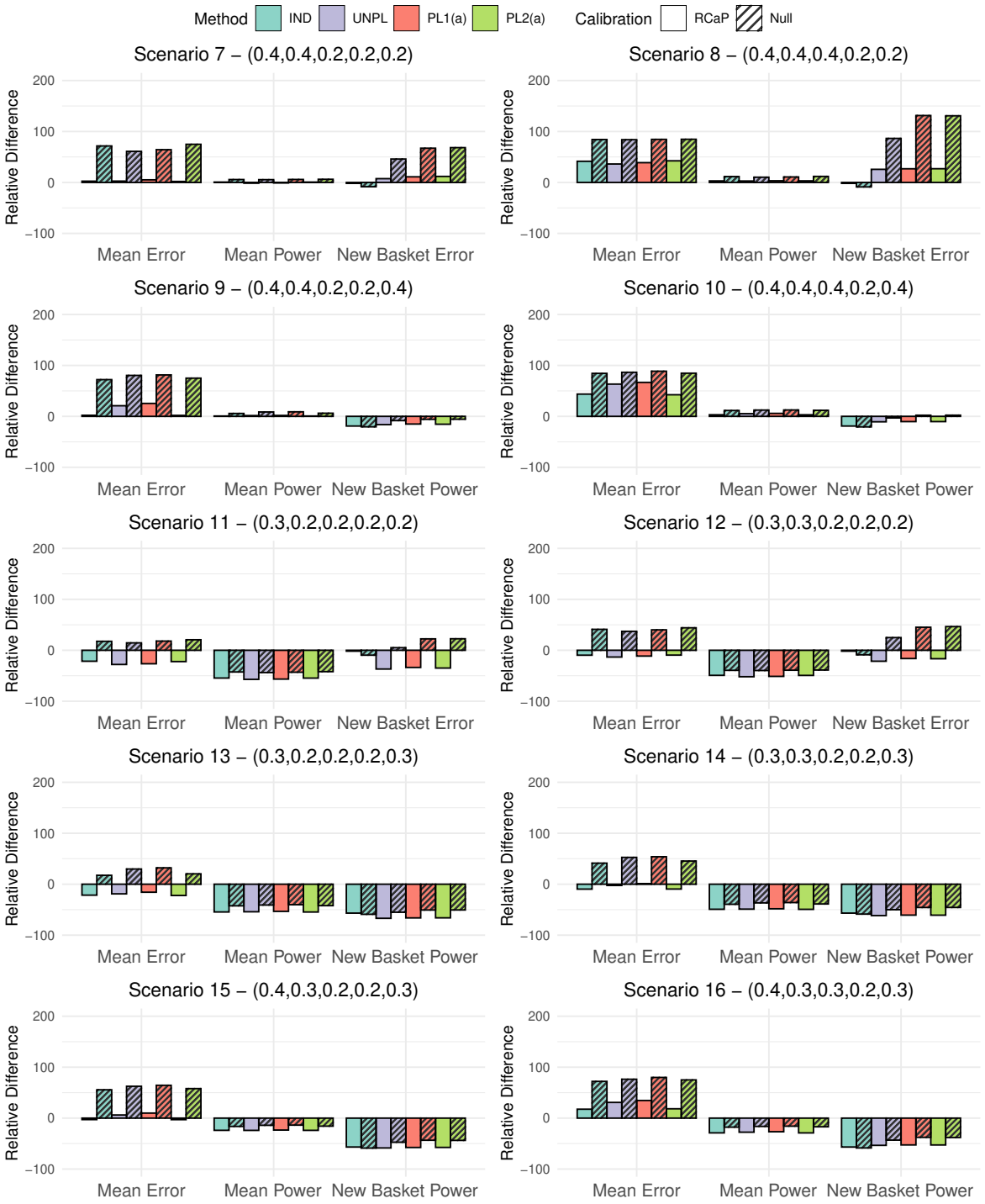


Figure B.1.1: The relative difference in type I error rate and power compared to the targeted values of 10% and 80% respectively. This is given for all four approaches for adding a basket under the two different calibration schemes, calibration under the global null and the RCaP. Results are split into 3 categories: mean error in which the percentage of data sets within which the null was rejected is averaged across all ineffective existing baskets; mean power as above but for all effective existing baskets and new basket error/power in which results are the percentage of data sets within which the null was rejected just in the new basket.

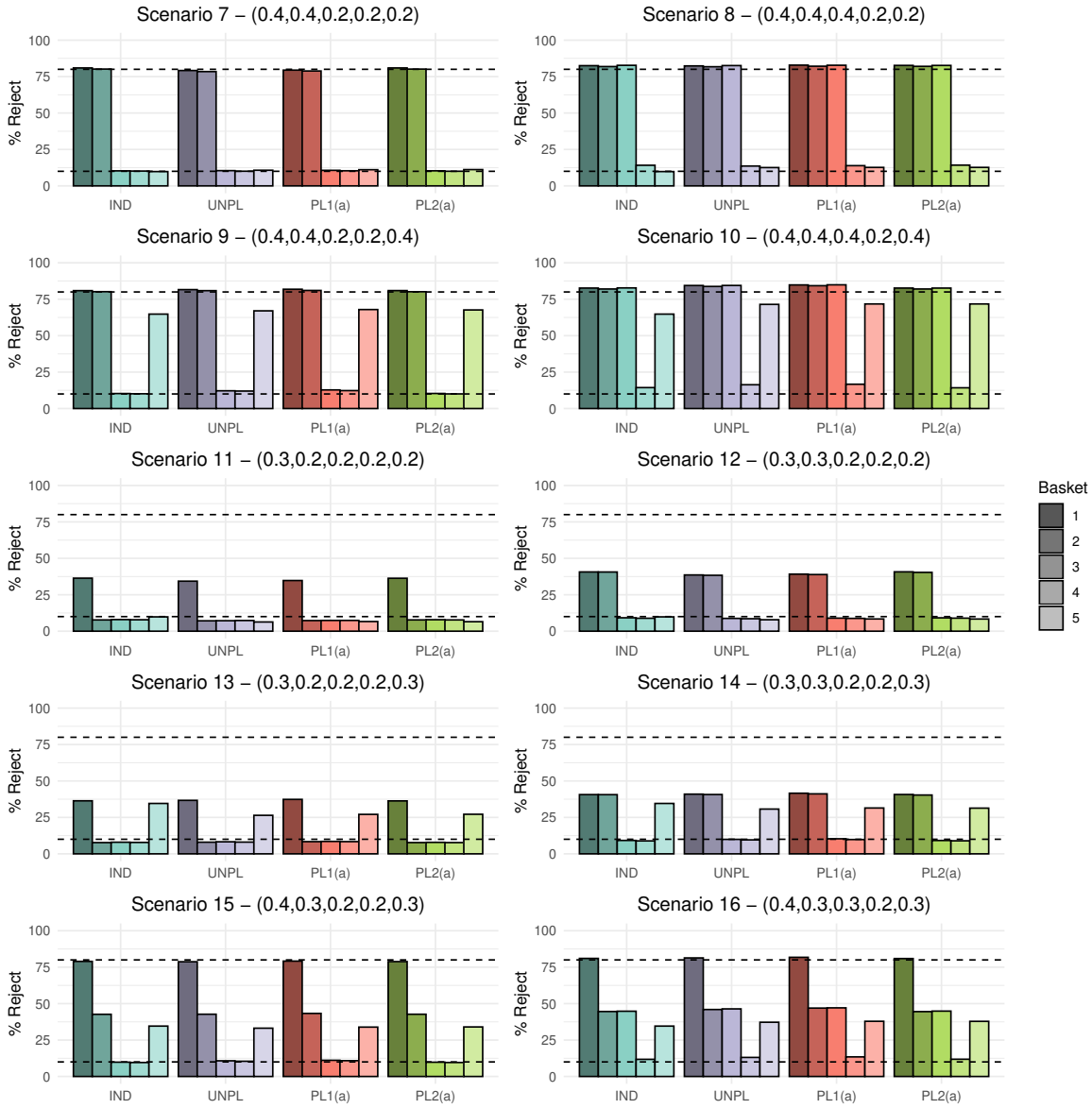


Figure B.1.2: Fixed scenario simulation study results: The percentage of data sets within which the null hypothesis was rejected, where Δ_{k_0} and $\Delta_{k'}$ were calibrated with RCaP to achieve a 10% type I error rate on average. This is plotted for each of the four approaches for adding a basket in all five baskets.

Table B.1.2: Operating characteristics for the fixed scenario simulation study in Chapter 3 under scenarios 1-4.

	% Reject					FWER	% Correct	Mean Point Estimate (Standard Deviation)				
Sc 1	**0.2**	**0.2**	**0.2**	**0.2**	**0.2**							
IND	6.33	6.52	6.42	6.46	9.82	29.37	70.63	0.202 (0.068)	0.202 (0.068)	0.202 (0.068)	0.203 (0.067)	0.200 (0.106)
UNPL	5.81	5.75	5.75	5.69	5.26	22.47	77.53	0.202 (0.065)	0.202 (0.066)	0.202 (0.065)	0.202 (0.064)	0.204 (0.079)
PL1(a)	5.73	5.92	5.89	5.78	5.45	22.82	77.18	0.202 (0.065)	0.202 (0.066)	0.202 (0.065)	0.202 (0.064)	0.204 (0.079)
PL2(a)	6.48	6.37	6.33	6.41	5.45	25.05	74.95	0.292 (0.068)	0.202 (0.068)	0.203 (0.068)	0.203 (0.067)	0.204 (0.079)
Sc 2	**0.4**	**0.2**	**0.2**	**0.2**	**0.2**							
IND	75.68	8.58	8.87	8.62	9.82	30.51	51.11	0.380 (0.096)	0.208 (0.072)	0.209 (0.071)	0.209 (0.070)	0.200 (0.106)
UNPL	73.73	7.83	8.12	8.01	7.07	25.47	52.69	0.376 (0.096)	0.208 (0.069)	0.209 (0.069)	0.209 (0.068)	0.212 (0.083)
PL1(a)	74.11	8.11	8.35	8.32	7.51	26.19	52.35	0.376 (0.096)	0.208 (0.069)	0.209 (0.069)	0.209 (0.068)	0.212 (0.083)
PL2(a)	75.67	8.49	8.70	8.62	7.38	27.58	52.91	0.380 (0.096)	0.208 (0.072)	0.209 (0.071)	0.209 (0.070)	0.212 (0.083)
Sc 3	**0.4**	**0.4**	**0.4**	**0.4**	**0.2**							
IND	86.74	86.06	86.86	86.85	9.82	9.82	54.18	0.399 (0.083)	0.398 (0.084)	0.399 (0.083)	0.399 (0.082)	0.200 (0.106)
UNPL	85.90	85.35	85.83	85.88	12.82	12.82	49.08	0.394 (0.082)	0.393 (0.083)	0.394 (0.082)	0.394 (0.081)	0.241 (0.096)
PL1(a)	86.45	85.92	86.12	86.42	13.00	13.00	50.33	0.394 (0.082)	0.393 (0.083)	0.394 (0.082)	0.394 (0.081)	0.241 (0.096)
PL2(a)	86.84	86.02	86.56	86.73	13.17	13.17	51.89	0.399 (0.083)	0.398 (0.084)	0.399 (0.083)	0.399 (0.082)	0.241 (0.096)
Sc 4	**0.4**	**0.4**	**0.4**	**0.4**	**0.4**							
IND	86.74	86.06	86.86	86.85	65.03	39.58		0.399 (0.083)	0.398 (0.084)	0.399 (0.083)	0.399 (0.082)	0.400 (0.131)
UNPL	88.57	88.12	88.53	88.51	72.25	47.45		0.399 (0.080)	0.399 (0.080)	0.399 (0.080)	0.400 (0.078)	0.398 (0.098)
PL1(a)	88.71	88.41	88.97	88.99	72.52	48.03		0.399 (0.080)	0.398 (0.080)	0.399 (0.079)	0.399 (0.078)	0.398 (0.098)
PL2(a)	86.84	86.02	86.56	86.73	72.46	44.33		0.399 (0.083)	0.398 (0.084)	0.399 (0.083)	0.399 (0.082)	0.398 (0.098)

Table B.1.3: Operating characteristics for the fixed scenario simulation study in Chapter 3 under scenarios 5-8.

	% Reject					FWER	% Correct	Mean Point Estimate (Standard Deviation)				
Sc 5	**0.2**	**0.2**	**0.2**	**0.2**	**0.4**							
IND	6.33	6.52	6.42	6.46	65.03	21.49	50.86	0.202 (0.068)	0.202 (0.068)	0.202 (0.068)	0.203 (0.067)	0.400 (0.131)
UNPL	7.16	7.28	7.51	7.31	53.41	24.19	38.22	0.207 (0.067)	0.206 (0.067)	0.207 (0.067)	0.207 (0.066)	0.365 (0.119)
PL1(a)	7.48	7.42	7.59	7.47	53.88	24.67	38.13	0.207 (0.067)	0.206 (0.067)	0.207 (0.067)	0.207 (0.066)	0.365 (0.119)
PL2(a)	6.48	6.37	6.33	6.41	53.84	21.29	40.94	0.202 (0.068)	0.202 (0.068)	0.203 (0.068)	0.203 (0.067)	0.365 (0.119)
Sc 6	**0.4**	**0.2**	**0.2**	**0.2**	**0.4**							
IND	75.68	8.58	8.87	8.62	65.03	22.89	36.84	0.380 (0.096)	0.208 (0.072)	0.209 (0.071)	0.209 (0.070)	0.400 (0.131)
UNPL	77.61	9.43	9.53	9.61	58.07	23.97	32.17	0.379 (0.093)	0.213 (0.071)	0.214 (0.071)	0.214 (0.070)	0.372 (0.115)
PL1(a)	77.73	9.75	9.75	9.75	59.16	24.28	33.31	0.379 (0.093)	0.213 (0.071)	0.214 (0.071)	0.214 (0.070)	0.372 (0.115)
PL2(a)	75.67	8.49	8.70	8.62	59.00	22.81	32.46	0.380 (0.096)	0.208 (0.072)	0.209 (0.071)	0.209 (0.070)	0.372 (0.115)
Sc 7	**0.4**	**0.4**	**0.2**	**0.2**	**0.2**							
IND	80.95	80.17	10.31	10.18	9.82	26.15	37.97	0.386 (0.092)	0.385 (0.093)	0.216 (0.075)	0.216 (0.074)	0.200 (0.106)
UNPL	79.17	78.44	10.46	10.06	10.74	26.14	44.87	0.381 (0.092)	0.380 (0.092)	0.216 (0.072)	0.216 (0.072)	0.221 (0.087)
PL1(a)	79.44	78.85	10.66	10.36	11.10	26.82	44.77	0.381 (0.092)	0.380 (0.092)	0.216 (0.072)	0.216 (0.072)	0.221 (0.087)
PL2(a)	80.95	80.15	10.34	10.04	11.18	27.06	46.64	0.386 (0.092)	0.385 (0.093)	0.216 (0.075)	0.216 (0.074)	0.221 (0.087)
Sc 8	**0.4**	**0.4**	**0.4**	**0.2**	**0.2**							
IND	82.50	81.89	82.74	14.15	9.82	22.64	41.66	0.392 (0.088)	0.391 (0.089)	0.392 (0.087)	0.223 (0.078)	0.200 (0.106)
UNPL	82.34	81.77	82.57	13.62	12.57	23.99	40.52	0.387 (0.087)	0.386 (0.088)	0.387 (0.087)	0.224 (0.076)	0.231 (0.092)
PL1(a)	82.86	82.13	82.78	13.88	12.68	24.32	40.75	0.387 (0.087)	0.386 (0.088)	0.387 (0.087)	0.224 (0.076)	0.231 (0.092)
PL2(a)	82.68	82.04	82.66	14.24	12.69	25.18	40.15	0.392 (0.088)	0.391 (0.088)	0.392 (0.087)	0.224 (0.078)	0.231 (0.092)

Table B.1.4: Operating characteristics for the fixed scenario simulation study in Chapter 3 under scenarios 9-12.

	% Reject					FWER	% Correct	Mean Point Estimate (Standard Deviation)				
Sc 9	**0.4**	**0.4**	**0.2**	**0.2**	**0.4**							
IND	80.91	80.19	10.3	10.12	64.80	18.17	34.09	0.385 (0.092)	0.385 (0.093)	0.216 (0.075)	0.216 (0.074)	0.399 (0.131)
UNPL	81.61	80.89	12.17	12.00	67.04	21.26	34.35	0.385 (0.089)	0.384 (0.089)	0.222 (0.075)	0.221 (0.074)	0.381 (0.109)
PL1(a)	81.86	81.04	12.78	12.3	67.93	22.05	34.41	0.385 (0.089)	0.384 (0.089)	0.222 (0.075)	0.222 (0.074)	0.381 (0.109)
PL2(a)	80.95	80.15	10.34	10.04	67.65	18.14	36.79	0.386 (0.092)	0.385 (0.093)	0.216 (0.075)	0.216 (0.074)	0.381 (0.109)
Sc 10	**0.4**	**0.4**	**0.4**	**0.2**	**0.4**							
IND	82.68	82.04	82.78	14.37	64.80	14.37	30.36	0.392 (0.088)	0.391 (0.089)	0.392 (0.087)	0.223 (0.078)	0.399 (0.131)
UNPL	84.43	83.79	84.46	16.33	71.47	16.33	36.58	0.393 (0.084)	0.392 (0.085)	0.392 (0.084)	0.228 (0.078)	0.390 (0.104)
PL1(a)	84.83	84.30	84.90	16.67	71.77	16.67	37.28	0.393 (0.084)	0.392 (0.085)	0.392 (0.084)	0.228 (0.078)	0.390 (0.104)
PL2(a)	82.68	82.04	82.66	14.24	71.77	14.24	33.97	0.392 (0.088)	0.391 (0.088)	0.392 (0.087)	0.224 (0.078)	0.390 (0.104)
Sc 11	**0.3**	**0.2**	**0.2**	**0.2**	**0.2**							
IND	36.38	7.73	7.98	7.84	9.83	28.40	23.71	0.287 (0.083)	0.206 (0.069)	0.207 (0.069)	0.207 (0.068)	0.200 (0.106)
UNPL	34.33	7.12	7.19	7.33	6.32	22.86	23.15	0.284 (0.081)	0.206 (0.067)	0.207 (0.066)	0.207 (0.065)	0.210 (0.080)
PL1(a)	34.77	7.29	7.38	7.39	6.63	23.42	23.22	0.284 (0.081)	0.206 (0.067)	0.207 (0.066)	0.207 (0.065)	0.210 (0.080)
PL2(a)	36.32	7.74	7.88	7.73	6.51	24.98	24.33	0.287 (0.083)	0.206 (0.069)	0.207 (0.069)	0.207 (0.068)	0.210 (0.080)
Sc 12	**0.3**	**0.3**	**0.2**	**0.2**	**0.2**							
IND	40.65	40.61	9.20	8.87	9.83	24.55	13.63	0.291 (0.081)	0.291 (0.082)	0.212 (0.070)	0.212 (0.069)	0.200 (0.106)
UNPL	38.59	38.36	8.78	8.56	7.85	21.50	12.14	0.288 (0.079)	0.287 (0.080)	0.212 (0.067)	0.212 (0.067)	0.215 (0.082)
PL1(a)	39.15	38.87	8.95	8.78	8.38	22.16	12.23	0.288 (0.079)	0.287 (0.080)	0.212 (0.067)	0.212 (0.067)	0.215 (0.081)
PL2(a)	40.72	40.36	9.16	8.94	8.34	22.83	13.34	0.291 (0.081)	0.290 (0.082)	0.212 (0.070)	0.212 (0.069)	0.215 (0.081)

Table B.1.5: Operating characteristics for the fixed scenario simulation study in Chapter 3 under scenarios 13-16.

	% Reject					FWER	% Correct	Mean Point Estimate (Standard Deviation)				
Sc 13	**0.3**	**0.2**	**0.2**	**0.2**	**0.3**							
IND	36.38	7.73	7.98	7.84	34.58	20.65	8.66	0.287 (0.083)	0.206 (0.069)	0.207 (0.069)	0.207 (0.068)	0.300 (0.122)
UNPL	36.70	8.00	8.26	8.05	26.43	20.68	7.47	0.287 (0.080)	0.209 (0.067)	0.210 (0.067)	0.210 (0.066)	0.285 (0.097)
PL1(a)	37.36	8.37	8.53	8.38	27.08	21.40	7.77	0.287 (0.080)	0.209 (0.067)	0.210 (0.067)	0.210 (0.066)	0.285 (0.097)
PL2(a)	36.32	7.74	7.88	7.73	27.16	20.54	7.18	0.287 (0.083)	0.206 (0.069)	0.207 (0.069)	0.207 (0.068)	0.285 (0.097)
Sc 14	**0.3**	**0.3**	**0.2**	**0.2**	**0.3**							
IND	40.65	40.61	9.20	8.87	34.58	16.34	5.08	0.291 (0.081)	0.291 (0.082)	0.212 (0.070)	0.212 (0.069)	0.300 (0.122)
UNPL	40.97	40.71	9.93	9.58	30.72	17.53	5.32	0.291 (0.078)	0.290 (0.079)	0.215 (0.068)	0.215 (0.067)	0.290 (0.095)
PL1(a)	41.60	41.16	10.33	9.82	31.45	17.96	5.62	0.291 (0.078)	0.290 (0.079)	0.215 (0.068)	0.215 (0.067)	0.290 (0.095)
PL2(a)	40.72	40.36	9.16	8.94	31.36	16.37	5.79	0.291 (0.081)	0.290 (0.082)	0.212 (0.070)	0.212 (0.069)	0.290 (0.095)
Sc 15	**0.4**	**0.3**	**0.2**	**0.2**	**0.3**							
IND	78.98	42.68	9.82	9.60	34.58	17.44	9.32	0.382 (0.093)	0.295 (0.083)	0.214 (0.072)	0.214 (0.072)	0.300 (0.122)
UNPL	78.74	42.73	10.77	10.45	33.14	18.92	9.94	0.378 (0.091)	0.295 (0.080)	0.217 (0.071)	0.217 (0.070)	0.295 (0.096)
PL1(a)	79.17	43.29	11.11	10.86	33.89	19.54	10.03	0.378 (0.091)	0.295 (0.080)	0.217 (0.071)	0.217 (0.070)	0.295 (0.096)
PL2(a)	78.84	42.72	9.79	9.61	34.00	17.46	10.98	0.382 (0.093)	0.295 (0.083)	0.214 (0.072)	0.214 (0.071)	0.295 (0.096)
Sc 16	**0.4**	**0.3**	**0.3**	**0.2**	**0.3**							
IND	81.03	44.59	44.79	11.75	34.58	11.75	4.65	0.383 (0.090)	0.299 (0.081)	0.300 (0.080)	0.219 (0.073)	0.300 (0.122)
UNPL	81.36	45.96	46.39	13.08	37.22	13.08	7.89	0.390 (0.088)	0.299 (0.078)	0.300 (0.077)	0.222 (0.071)	0.301 (0.094)
PL1(a)	81.78	46.94	47.1	13.46	37.88	13.46	8.55	0.380 (0.088)	0.300 (0.078)	0.300 (0.077)	0.222 (0.071)	0.301 (0.094)
PL2(a)	80.93	44.52	44.87	11.83	37.84	11.83	5.69	0.383 (0.091)	0.299 (0.081)	0.300 (0.080)	0.219 (0.073)	0.301 (0.094)

B.2 Comparison of Using Differing Number of Scenarios in the RCaP

Simulation studies in Chapter 3 were conducted under the novel robust calibration procedure (RCaP) in order to achieve a 10% type I error rate on average across several scenarios. RCaP was implemented under scenarios 1, 2, 3, 7 and 8 in Chapter 3. These scenarios included all global and partial nulls assuming equal sample sizes across baskets. However, due to the new basket having a reduced sample size, these scenarios no longer cover all partial and global nulls. This is resolved by also including scenarios 5, 6, 9 and 10 in the calibration procedure. Exploration is now conducted into differences in performance based on the number of scenarios incorporated into RCaP.

Note that under UNPL, calibration differs as it consists of just the four existing baskets. The equal sample size across baskets, results in just 4 global and partial null scenarios and thus Δ_{k_0} is calibrated just across these four scenarios. Results presented incorporate the irrelevant difference between calibration in UNPL, with absolute difference values given as 0 throughout.

For all approaches, $\Delta_{k'}$ values are equal under both calibrations. Under scenarios 5, 6, 9 and 10, the baskets response rate is effective and thus not included when taking the quantile to obtain $\Delta_{k\prime}$, therefore, only including scenarios 1, 2, 3, 7 and 8 in the calibration.

Figure B.2.1 presents the absolute difference in percentage rejections of the null under an RCaP under scenarios 1, 2, 3, 7 and 8 and an RCaP under scenarios 1-10 (excluding the global alternative). In all bar a handful of cases, the percentage rejections i.e. type I error rate and power are lower under an RCaP under 1-10 vs. the RCaP with fewer scenarios.

Differences under IND and PL2 are always less than 1%, a negligible difference. This is expected due to the very similar Δ_{k_0} and $\Delta_{k\prime}$ values obtained under both calibration

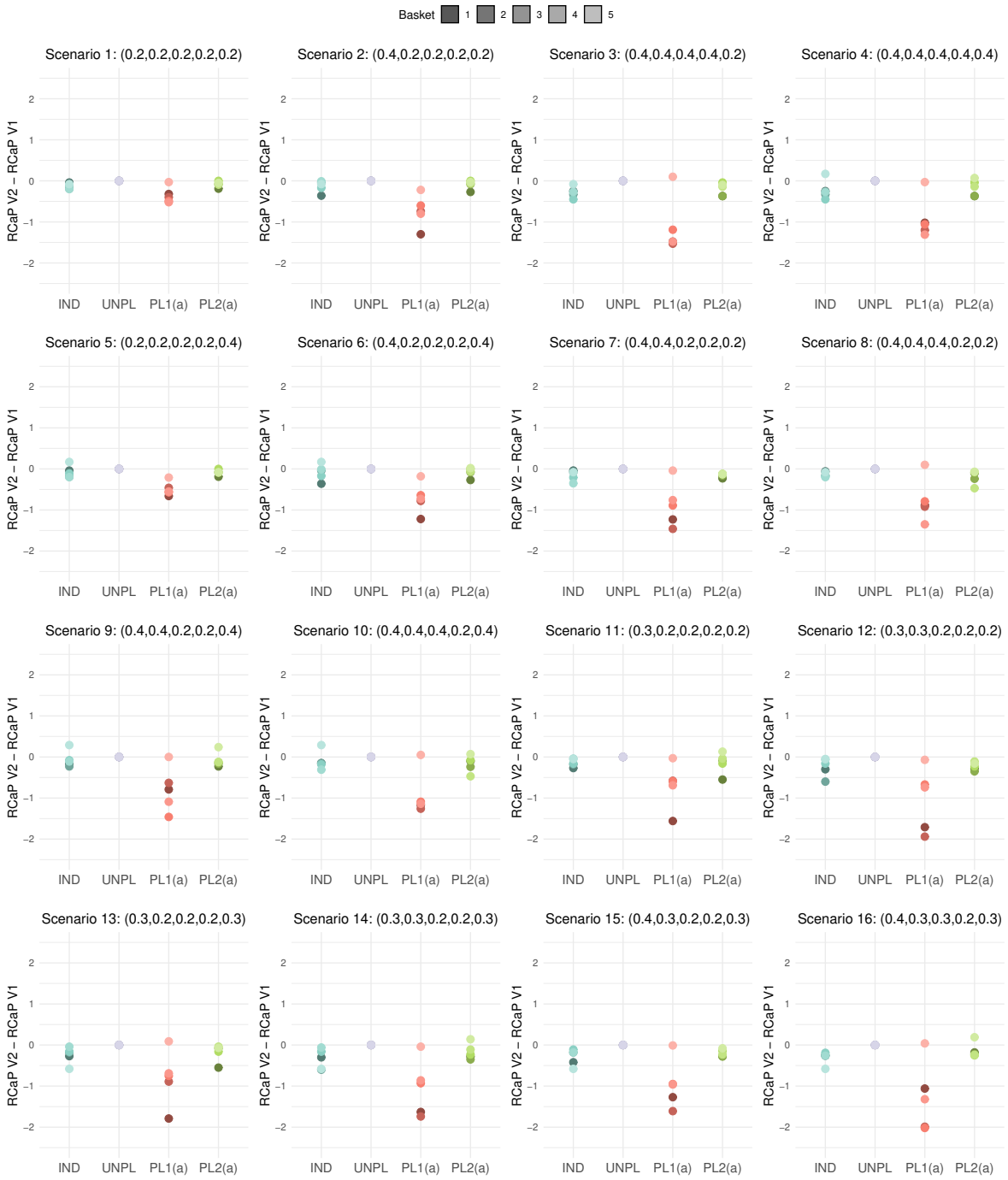


Figure B.2.1: Absolute difference in the number of simulated data sets within which the null hypothesis is rejected between an RCaP under scenarios 1, 2, 3, 7 and 8 (RCaP V1) and an RCaP under scenarios 1-10 (RCaP V2), excluding the global alternative. This is split by approach and basket.

Table B.2.1: Calibrated Δ_{k_0} and $\Delta_{k\prime}$ values for each of the approaches for adding a basket under an RCaP under scenarios 1, 2, 3, 7 and 8 and an RCaP under scenarios 1-10.

	RCaP across 1-10		RCaP across 1,2,3,7,8	
	Δ_{k_0}	$\Delta_{k\prime}$	Δ_{k_0}	$\Delta_{k\prime}$
IND	0.9044	0.8989	0.9030	0.8989
UNPL	0.9056	0.9056	0.9056	0.9056
PL1(a)	0.9101	0.9021	0.9034	0.9021
PL2(a)	0.9044	0.9021	0.9030	0.9021

cases (see Table B.2.1). However, more differences are observed under PL1(a), reaching up to 2% (scenario 16 in a marginally effective basket). This is due to the more conservative cut-off value. So even in the worst cases, differences between approaches are rather small. Calibration across fewer scenarios is less computationally expensive as it considers four fewer data scenarios compared to a calibration across scenarios 1-10 (excluding scenario 4, the global alternative). Due to the very minute differences between approaches, particularly in IND and PL2(a), a calibration across scenarios 1, 2, 3, 7 and 8 is recommended for it's reduced computational time.

B.3 Simulation Results Using Different Scenario Weights in the RCaP

In all simulation studies presented in Chapter 3, for the RCaP equal weights, ω_i, for each of the i scenarios are implemented in the procedure. Equal weights implies each of the scenarios carries the same importance in the calibration, this may be the case if each scenario is equally likely to occur in the trial. This section explores the effect of altering these weights on the performance of the calibration procedure.

The simulation studies presented all implemented RCaP across the 5 scenarios presented in Table B.3.1, with equal weights $\omega_1 = \omega_2 = \omega_3 = \omega_4 = \omega_5 = 1$ implemented. These weights are now varied to put more importance on certain scenarios relative to

Table B.3.1: Simulation study scenarios included in the RCaP in Chapter 3.

	p_1	p_2	p_3	p_4	p_5
Scenario 1	0.2	0.2	0.2	0.2	0.2
Scenario 2	0.4	0.2	0.2	0.2	0.2
Scenario 3	0.4	0.4	0.2	0.2	0.2
Scenario 4	0.4	0.4	0.4	0.2	0.2
Scenario 5	0.4	0.4	0.4	0.4	0.2

others. Tables B.3.2, B.3.3, B.3.4 and B.3.5 summarise the operating characteristics of each of the four approaches for adding: IND, UNPL, PL1 and PL2, under different weight settings. Presented are the calibrated cut-off values obtained and the mean type I error rate and power (split by new and existing baskets). Note that the mean is taken across scenarios 1-10 presented in Table B.1.1 in which the basket has either an effective or ineffective response rate.

Consider first the IND approach for adding a basket. As displayed in Table B.3.2, cut-off values, mean error and mean power are identical for the new basket across all weight combinations. Under the IND approach, new baskets are analysed as independent, which guarantees error control to the nominal level in the new basket in all scenarios in which the true response rate is q_0. As such, the cut-off value obtained under each of the 5 scenarios under considered will be equal, so altering the weight will have no impact. However, operating characteristics in the existing baskets are affected by the weight choice. As mentioned in the discussion of Chapter 3, the type I error rate increases with the number of effective existing baskets, thus scenarios 3 and 4 will display greater error rates than say scenarios 1 and 2. Placing more weight on the scenarios with only 2 or 3 ineffective baskets (i.e. where error inflation is expected to be the greatest) results in a more conservative cut-off value, Δ_{k_0}, in order to ensure error control. With this a reduction in power is observed compared to equal weights. Under equal weights, the mean power is 83.2%, whereas placing double the weight on scenario 4 results in a power of 82.1%. Placing 4 times the weight on this scenario

decreases the power further to 81.2%. If more weight is placed on scenario 1 (where the type I error rate is expected to be lowest due to all baskets being null), Δ_{k_0} is less conservative than equal weights, resulting in a higher mean error of 9.4% compared to 8.8% but with an increase in power of 84.2%.

Table B.3.2: IND: Summary of operating characteristics under several weight combinations ($\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$) for the 5 scenarios included in the RCaP in Chapter 3.

IND			Mean Error		Mean Power	
$\boldsymbol{\omega}$	Δ_{k_0}	$\Delta_{k'}$	Existing	New	Existing	New
(1,1,1,1,1)	0.902	0.899	8.75	9.97	83.18	65.19
(2,1,1,1,1)	0.896	0.899	9.39	9.97	84.15	65.19
(1,2,1,1,1)	0.900	0.899	8.99	9.97	83.54	65.19
(1,1,2,1,1)	0.902	0.899	8.73	9.97	83.15	65.19
(1,1,1,2,1)	0.908	0.899	8.12	9.97	82.12	65.19
(1,1,1,1,2)	0.902	0.899	8.75	9.97	83.18	65.19
(1,1,1,1,4)	0.902	0.899	8.75	9.97	83.18	65.19
(1,1,1,4,1)	0.913	0.899	7.64	9.97	81.20	65.19
(1,1,1,2,2)	0.908	0.899	8.12	9.97	82.12	65.19

Under an unplanned addition, the calibrated cut-off values for the UNPL approach do vary based on the weights implemented in RCaP for all baskets existing and new. Like in the IND approach, placing more weights on scenarios 1 and 2 gives less conservative cut-off values for all baskets resulting in higher error with higher power compared to equal weights. Similarly, placing more weight on scenarios with fewer ineffective baskets requires more conservative cut-off values to ensure error control with a lower power also observed. As cut-off values are calibrated based on just the existing baskets, any scenarios which put equal weight on existing baskets will be equivalent to the $\boldsymbol{\omega} = (1, 1, 1, 1, 1)$ case, regardless of the choice of ω_5.

PL1(a) borrows information between all baskets therefore, changing the weights in all scenarios will result in differing operating characteristics. Similar findings in terms of conservative calibrated cut-offs to the IND and UNPL approach are drawn. When ω_5 is increased relative to the weights on other scenarios, the cut-off value is again more conservative than an equal weight scenario, particularly for the new basket.

Table B.3.3: UNPL: Summary of operating characteristics under several weight combinations ($\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$) for the 5 scenarios included in the RCaP in Chapter 3.

UNPL			Mean Error		Mean Power	
$\boldsymbol{\omega}$	Δ_{k_0}	$\Delta_{k'}$	Existing	New	Existing	New
(1,1,1,1,1)	0.900	0.900	9.63	10.09	84.33	65.38
(2,1,1,1,1)	0.893	0.893	10.40	10.88	85.30	66.62
(1,2,1,1,1)	0.897	0.897	9.89	10.34	84.69	65.80
(1,1,2,1,1)	0.901	0.901	9.47	9.99	84.13	65.16
(1,1,1,2,1)	0.907	0.907	8.87	9.59	83.23	64.32
(1,1,1,1,2)	0.900	0.900	9.63	10.09	84.33	65.38
(1,1,1,1,4)	0.900	0.900	9.63	10.09	84.33	65.38
(1,1,1,4,1)	0.912	0.912	8.23	9.22	82.20	63.40
(1,1,1,2,2)	0.907	0.907	8.87	9.59	83.23	64.32

In scenarios 5, only the new basket is ineffective, thus this scenario only contributes to the calibration of $\Delta_{k'}$ and does not impact Δ_{k_0}. Under $\boldsymbol{\omega} = (1, 1, 1, 1, 2)$, $\Delta_{k'}$ increases to 0.909 compared to 0.901 under equal weights and further increases to 0.923 under $\boldsymbol{\omega} = (1, 1, 1, 1, 4)$, resulting in a lower power in the new basket of 61.2% compared to 65.17% under equal weights. The most conservative Δ_{k_0} is observed under $\boldsymbol{\omega} = (1, 1, 1, 4, 1)$ in which $\Delta_{k_0} = 0.912$ resulting in 82.2% power in existing baskets compared to 84.2% under equal weights. Identical conclusions are drawn for the PL2(a) approach as displayed in Table B.3.5.

To summarise, weights do play an important role in the RCaP procedure and can be utilised in order to influence error control and power improvement. As seen, placing more weight on scenarios with fewer ineffective baskets will improve error control with a cost of reduced power, whilst putting more weight on scenarios with mostly ineffective baskets gives better power. Should information be available regarding which scenarios are most likely to occur, these weights could be specified in order to improve trial inference.

Table B.3.4: PL1(a): Summary of operating characteristics under several weight combinations ($\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$) for the 5 scenarios included in the RCaP in Chapter 3.

PL1(a)			Mean Error		Mean Power	
$\boldsymbol{\omega}$	Δ_{k_0}	$\Delta_{k'}$	Existing	New	Existing	New
(1,1,1,1,1)	0.900	0.901	9.57	9.99	84.22	65.17
(2,1,1,1,1)	0.894	0.894	10.37	10.83	85.24	66.48
(1,2,1,1,1)	0.898	0.897	9.87	10.42	84.62	65.84
(1,1,2,1,1)	0.901	0.903	9.46	9.80	84.08	64.82
(1,1,1,2,1)	0.907	0.908	8.51	9.50	83.19	64.08
(1,1,1,1,2)	0.901	0.909	9.57	9.45	84.22	63.95
(1,1,1,1,4)	0.901	0.923	9.57	8.41	84.22	61.20
(1,1,1,4,1)	0.912	0.919	8.19	8.76	82.20	62.09
(1,1,1,2,2)	0.907	0.916	8.85	9.00	83.29	62.73

Table B.3.5: PL2(a): Summary of operating characteristics under several weight combinations ($\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \omega_4, \omega_5)$) for the 5 scenarios included in the RCaP in Chapter 3.

PL2(a)			Mean Error		Mean Power	
$\boldsymbol{\omega}$	Δ_{k_0}	$\Delta_{k'}$	Existing	New	Existing	New
(1,1,1,1,1)	0.902	0.901	8.71	10.00	83.14	65.17
(2,1,1,1,1)	0.896	0.893	9.36	10.83	84.11	66.60
(1,2,1,1,1)	0.900	0.897	8.96	10.41	83.54	65.82
(1,1,2,1,1)	0.903	0.903	8.67	9.81	83.09	64.88
(1,1,1,2,1)	0.908	0.908	8.15	9.49	82.14	64.02
(1,1,1,1,2)	0.902	0.909	8.71	9.44	83.14	63.85
(1,1,1,1,4)	0.902	0.924	8.71	8.40	83.14	62.74
(1,1,1,4,1)	0.913	0.919	7.63	8.76	81.22	62.08
(1,1,1,2,2)	0.908	0.916	8.15	9.00	82.14	62.74

B.4 Fixed Scenario Simulation Results For Calibration Under the Global Null

Results presented in Chapter 3 utilised the RCaP in which the type I error rate is controlled on average across several data scenarios. The results under a traditional calibration approach in which type I error rate is controlled under a global null scenario, are presented here.

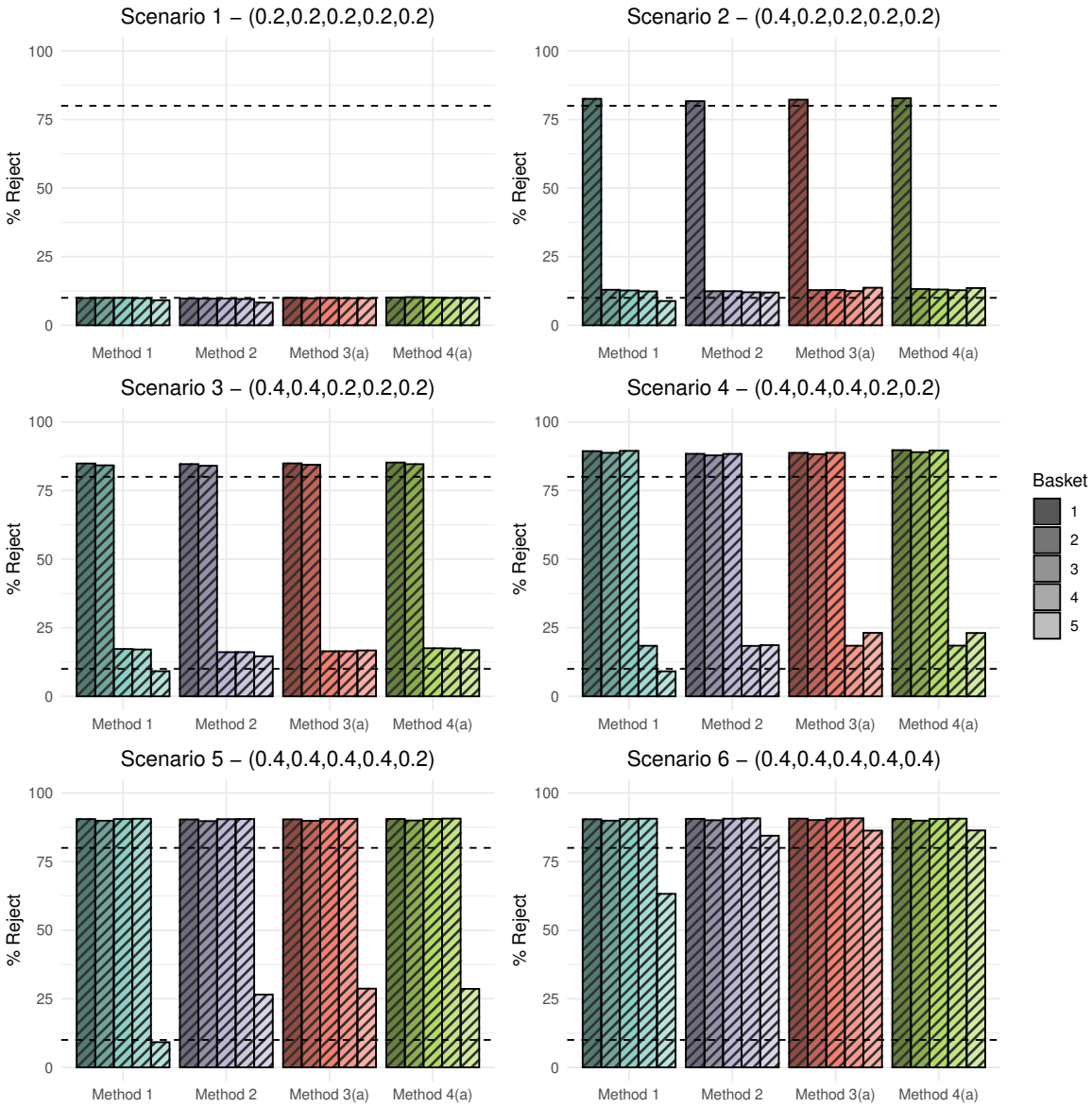


Figure B.4.1: Fixed scenario simulation study results: The percentage of data sets within which the null hypothesis was rejected, where Δ was calibrated under the null to achieve a 10% type I error rate on average. This is plotted for each of the four approaches for adding a basket for all five baskets.

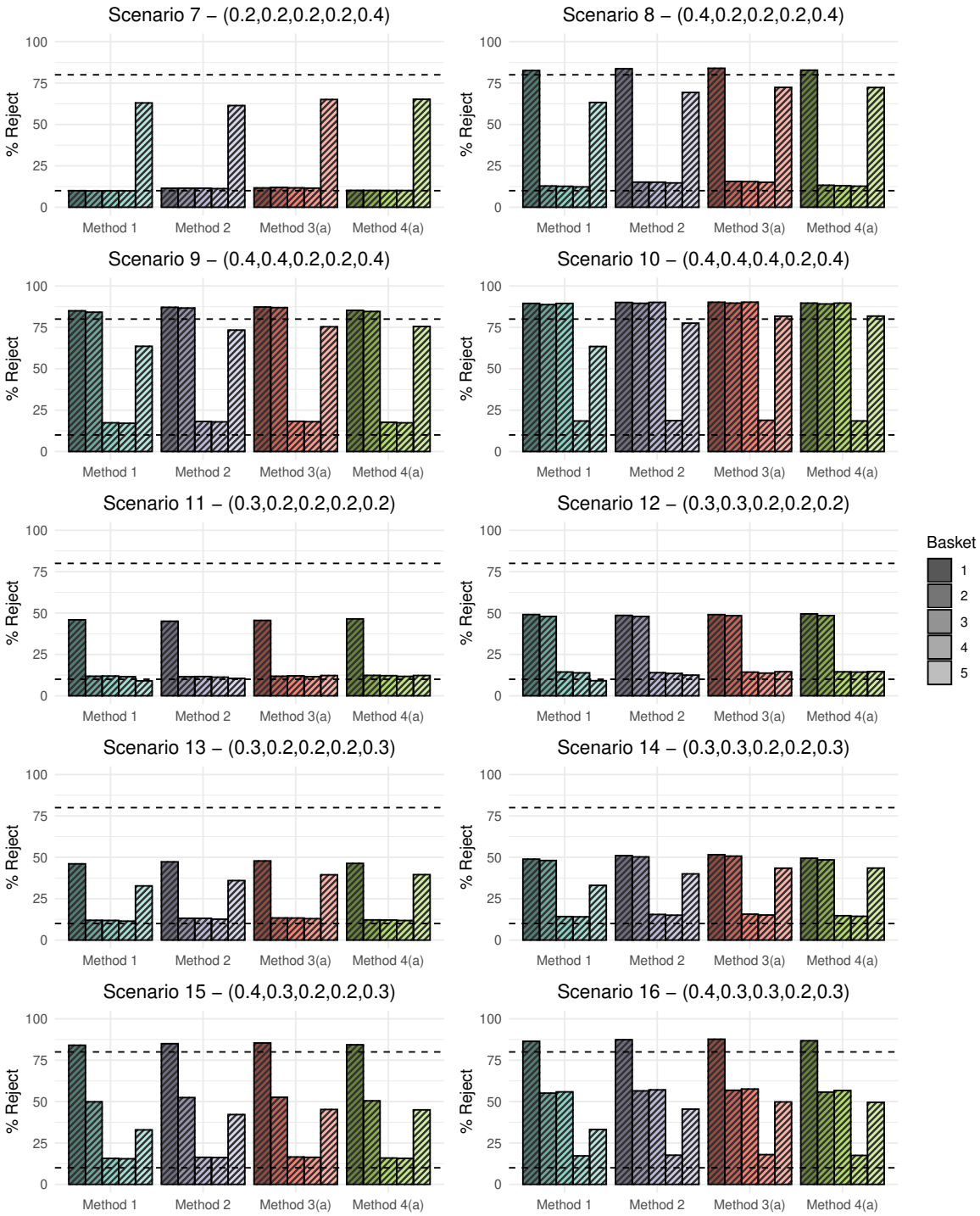


Figure B.4.2: Fixed scenario simulation study results: The percentage of data sets within which the null hypothesis was rejected, where Δ was calibrated under the null to achieve a 10% type I error rate on average. This is plotted for each of the four approaches for adding a basket for all five baskets.

B.5 Random Scenario Simulation

Results of pair-wise comparisons between approaches for the simulation study presented in Section 3.3.5 of Chapter 3. Figures B.5.1 and B.5.2 present these pair-wise comparisons split into existing and new baskets respectively. Tables B.5.1 and B.5.2 display full results for all 12 simulation study settings.

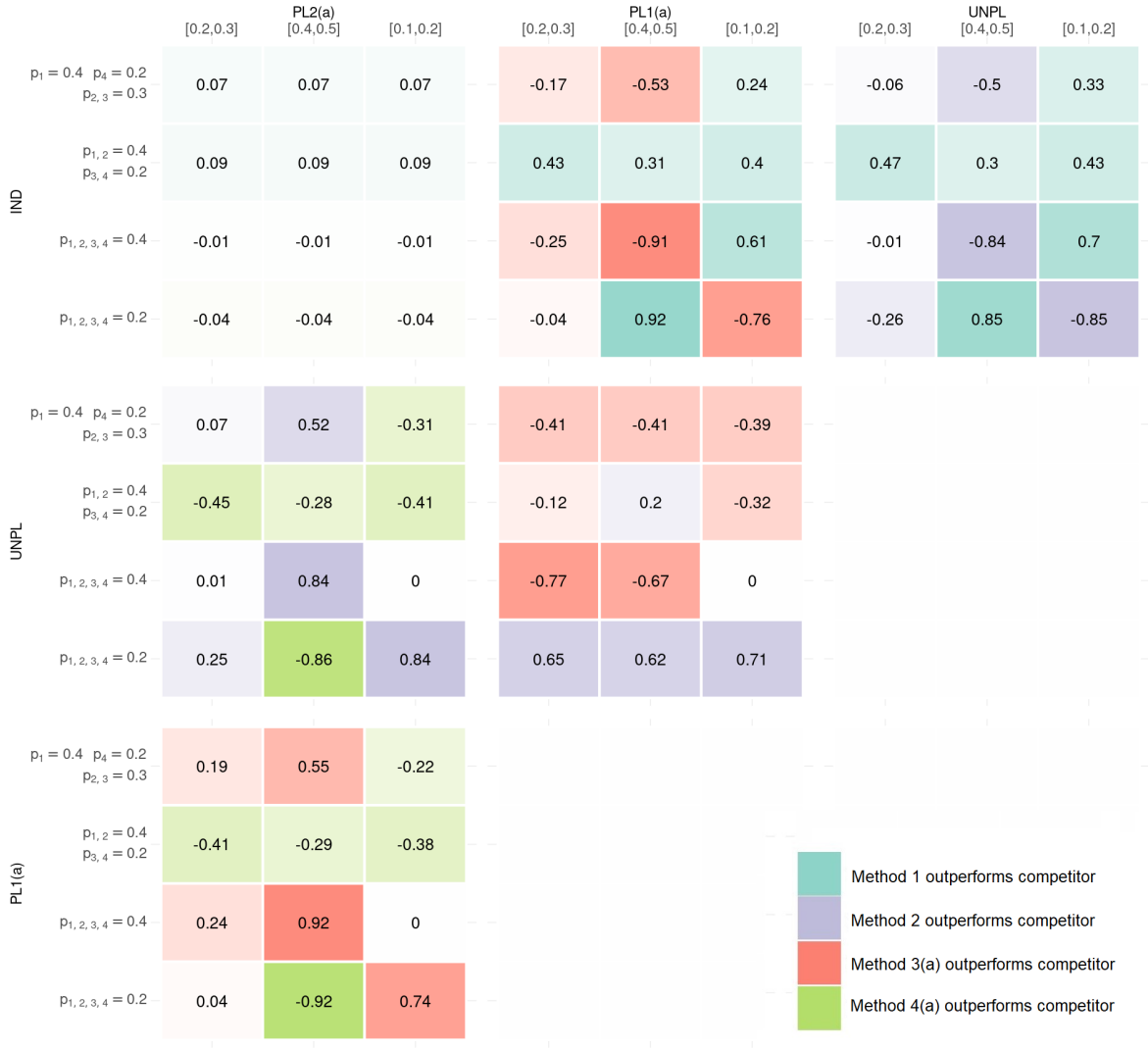


Figure B.5.1: Pair-wise comparison between approaches in each of the 12 simulation settings within which the true response rate in the new basket is varied. The heat map presents the difference in proportion of times the approach corresponding to row gave a correct conclusion over the approach corresponding to column when discrepancies between the two approaches arise in existing baskets only.

First consider just existing baskets and the pair-wise comparisons. Note that IND and PL2(a) are equivalent in these baskets, hence values in such a comparison are centred around 0 with slight simulation error. Results for other comparisons are akin to those presented in Chapter 3, indicating that the driving force behind the results presented in Chapter 3 are the difference in proportion of correct conclusion when discrepancies lie in existing baskets.

Then looking at pair-wise comparisons in just the singular new basket, in this case PL1(a) and PL2(a) are equivalent and so results are centred around 0 but with rather a lot of simulation noise. In the comparison between IND and UNPL, some cases result in all correct conclusions occurring for just one of the two approaches in discrepancies. For example, in the case where homogeneity between new and existing baskets with all having a null response rate, UNPL in which information is borrowed between all baskets leads to correct conclusions in all 309 cases of discrepancies. Whilst in the case of heterogeneity when the existing baskets are effective with the new basket ineffective, IND where the new is analysed independently leads to the correct conclusion in all 40 discrepancies. The number of cases where UNPL outperforms IND differs when looking at just the new basket compared to all discrepancies, with simulations in which the new basket is ineffective now often preferring UNPL.

Much more substantial differences are observed in the comparison between UNPL and PL1(a) under just the new basket compared to overall discrepancies. Previously, in all cases bar when the existing baskets are all null, PL1(a) outperformed UNPL, i.e. a planned addition is preferred to unplanned. However, when considering just the new basket this reverses with UNPL now only preferred when the new basket is ineffective. This arises from the more conservative $\Delta_{k\prime}$ cut-off under UNPL compared to PL1(a). Note that in most cases very few discrepancies between conclusions under both approaches arise. For instance, when all existing baskets are effective no discrepancies arise when the new basket is ineffective and only 1 or 2 discrepancies arising when it is

either marginally effective or effective.

Similar comparisons between UNPL and PL2(a) can be drawn as between UNPL and PL1(a) with cases in which UNPL outperforms PL1(a) also resulting in a conclusions that UNPL outperformed PL2(a). Pair-wise comparisons between IND and PL1(a) approaches and IND and PL2(a) approaches result in the same conclusions as those made in Chapter 3.

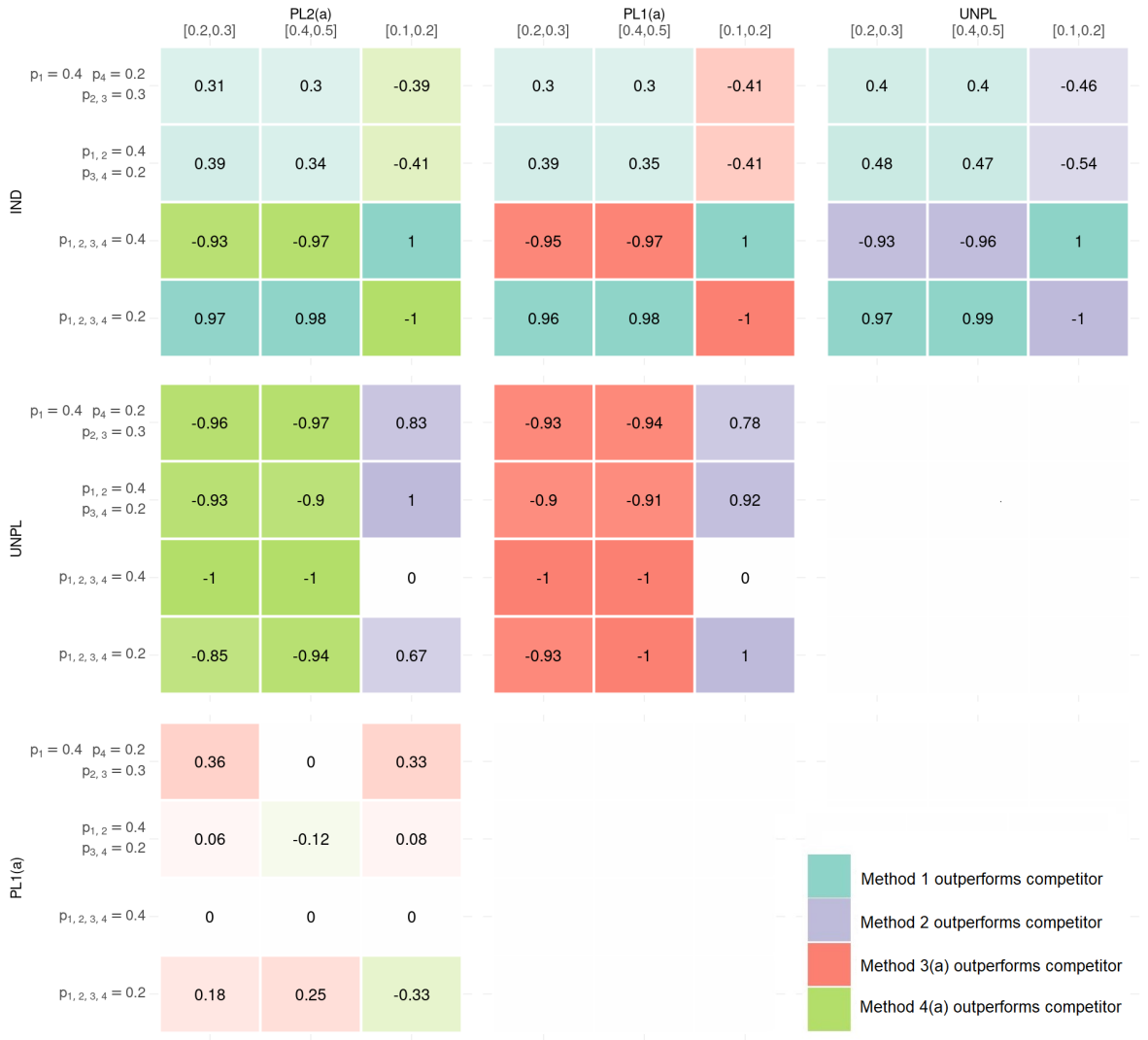


Figure B.5.2: Pair-wise comparison between approaches in each of the 12 simulation settings within which the true response rate in the new basket is varied. The heat map presents the difference in proportion of times the approach corresponding to row gave a correct conclusion over the approach corresponding to column when discrepancies between the two approaches arise in the new basket only.

Table B.5.1: Overall error rates and power for the varied truth simulation study in which the truth in the new basket is varied with the response rate in existing baskets fixed under settings 1 and 2.

	% Reject					FWER	% All Correct
Setting 1(a)	**0.2**	**0.2**	**0.2**	**0.2**	**[0.2,0.3]**		
IND	6.40	6.39	6.29	6.35	24.33	21.09	19.17
UNPL	5.94	6.09	5.99	6.19	12.97	19.96	8.79
PL1(a)	6.16	6.39	6.23	6.44	13.23	20.83	8.76
PL2(a)	6.36	6.41	6.29	6.33	13.19	21.12	9.49
Setting 1(b)	**0.2**	**0.2**	**0.2**	**0.2**	**[0.4,0.5]**		
IND	6.40	6.39	6.29	6.35	80.73	21.08	63.45
UNPL	7.45	7.33	7.57	7.43	66.96	24.76	48.18
PL1(a)	7.67	7.52	7.77	7.65	67.33	25.43	47.88
PL2(a)	6.36	6.41	6.29	6.33	67.29	21.12	51.67
Setting 1(c)	**0.2**	**0.2**	**0.2**	**0.2**	**[0.1,0.2]**		
IND	6.40	6.39	6.29	6.35	5.11	24.95	75.05
UNPL	5.45	5.34	5.29	5.33	2.02	19.05	80.95
PL1(a)	5.58	5.60	5.57	5.50	2.07	19.76	80.24
PL2(a)	6.36	6.41	6.29	6.33	2.06	22.41	77.59
Setting 2(a)	**0.4**	**0.4**	**0.4**	**0.4**	**[0.2,0.3]**		
IND	88.86	86.00	86.81	86.87	24.33	0.02	15.24
UNPL	86.74	86.17	86.85	86.91	25.75	0.03	17.35
PL1(a)	87.24	86.82	87.44	87.54	25.76	0.03	17.51
PL2(a)	86.86	86.02	86.79	86.90	25.76	0.03	16.03
Setting 2(b)	**0.4**	**0.4**	**0.4**	**0.4**	**[0.4,0.5]**		
IND	86.86	86.00	86.81	86.87	80.72	0.00	49.60
UNPL	88.73	88.26	88.92	88.85	82.49	0.00	53.85
PL1(a)	89.10	88.64	89.16	89.31	82.51	0.00	54.27
PL2(a)	86.86	86.02	86.79	86.90	82.51	0.00	50.58
Setting 2(c)	**0.4**	**0.4**	**0.4**	**0.4**	**[0.1,0.2]**		
IND	86.86	86.00	86.81	86.87	5.11	5.11	58.09
UNPL	84.50	83.80	84.59	84.65	5.51	5.51	50.19
PL1(a)	84.86	84.33	85.02	85.21	5.51	5.51	51.76
PL2(a)	86.86	86.02	86.79	86.90	5.51	5.51	57.86

Table B.5.2: Overall error rates and power for the varied truth simulation study in which the truth in the new basket is varied with the response rate in existing baskets fixed under settings 3 and 4.

			% Reject			FWER	% All Correct
Setting 3(a)	**0.4**	**0.4**	**0.2**	**0.2**	**[0.2,0.3]**		
IND	80.88	80.16	10.10	9.88	24.33	17.61	13.19
UNPL	80.02	79.26	10.79	10.48	22.24	18.82	12.08
PL1(a)	80.34	79.75	11.08	10.76	22.81	19.29	12.30
PL2(a)	80.87	80.15	10.20	9.88	22.79	17.68	13.37
Setting 3(b)	**0.4**	**0.4**	**0.2**	**0.2**	**[0.4,0.5]**		
IND	80.88	80.16	10.10	9.88	80.72	17.59	43.16
UNPL	81.75	81.06	11.92	11.79	78.25	20.75	40.82
PL1(a)	82.04	81.30	12.35	12.32	79.08	21.59	40.72
PL2(a)	80.87	80.15	10.20	9.88	29.15	17.66	43.58
Setting 3(c)	**0.4**	**0.4**	**0.2**	**0.2**	**[0.1,0.2]**		
IND	80.88	80.16	10.10	9.88	5.11	21.81	50.96
UNPL	78.84	77.95	9.62	9.25	4.47	20.55	48.73
PL1(a)	79.40	78.47	9.80	9.50	4.70	21.01	49.13
PL2(a)	80.87	80.15	10.20	9.88	4.71	21.44	50.91
Setting 4(a)	**0.4**	**0.3**	**0.3**	**0.2**	**[0.2,0.3]**		
IND	80.	86.00	86.81	86.87	24.33	11.76	3.44
UNPL	80.80	44.94	44.94	12.18	22.76	12.21	4.89
PL1(a)	81.18	45.38	45.52	12.52	23.32	12.55	5.12
PL2(a)	80.77	44.42	44.57	11.89	23.22	11.92	3.69
Setting 4(b)	**0.4**	**0.3**	**0.3**	**0.2**	**[0.4,0.5]**		
IND	80.91	44.39	44.46	11.74	80.72	11.74	10.95
UNPL	82.42	47.06	47.63	13.90	78.88	13.90	14.40
PL1(a)	82.81	47.66	48.22	14.26	79.46	14.26	15.13
PL2(a)	80.77	44.42	44.57	11.89	23.22	11.92	3.69
Setting 4(c)	**0.4**	**0.3**	**0.3**	**0.2**	**[0.1,0.2]**		
IND	80.91	44.39	44.46	11.74	80.72	11.74	10.95
UNPL	82.42	47.06	47.63	13.90	78.88	13.90	14.40
PL1(a)	82.81	47.66	48.22	14.26	79.46	14.26	15.13
PL2(a)	80.77	44.42	44.57	11.89	79.46	11.89	11.11

B.6 Investigating the Robustness to the Timing of Addition

In previous simulation studies it is assumed that the timing of addition of a new basket is known prior to the trial, however, this could easily not be the case. This section explores the effect of timing of addition on the performance of analysis methods. Timing of addition is explored by varying the sample size in the new basket. Baskets added early in the trial have a larger sample size than those added at a later time point.

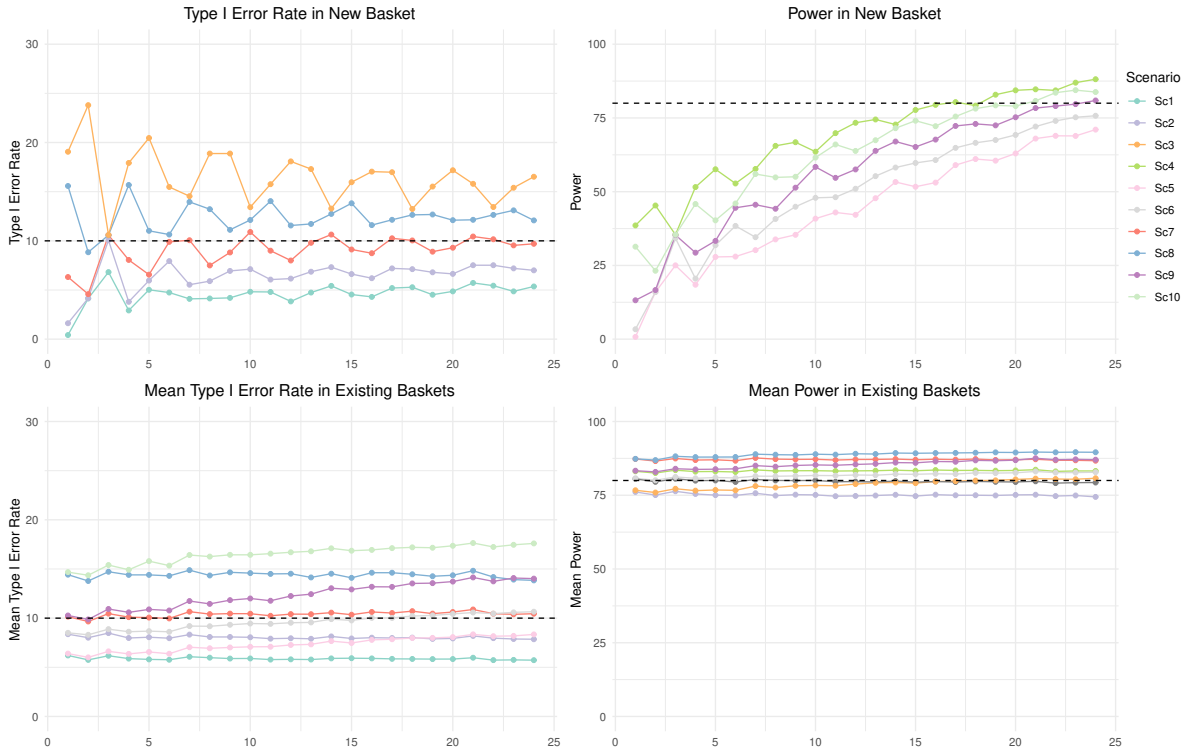


Figure B.6.1: Type I error rate and power under each sample size of n_5 from 1 to 24 by applying PL1(b), split by existing and new baskets.

When a basket is added later in the trial sample sizes may be smaller. This small basket sample size will result in a lack of power and precision of treatment effect estimates. Borrowing information from the larger existing baskets will prove more beneficial in such a setting compared to when the new basket is larger and of similar size to existing baskets. Thus, methods that utilise information borrowing in the new basket

become increasingly superior over the IND approach in which an independent analysis is conducted. Improvements in performance in the new basket can still be obtained via information borrowing regardless of timing of addition, however, it is those added later in the trial that will benefit more substantially due to their reduced sample size. Both PL1(a) and PL2(a) make a planned addition of a basket whilst utilising information borrowing so timing of addition is taken into account in both the calibration process and analysis. In the case that the sample size of the new baskets is unknown, performance of these approaches is more liable to change and thus the robustness of the approaches to the timing of addition is now explored.

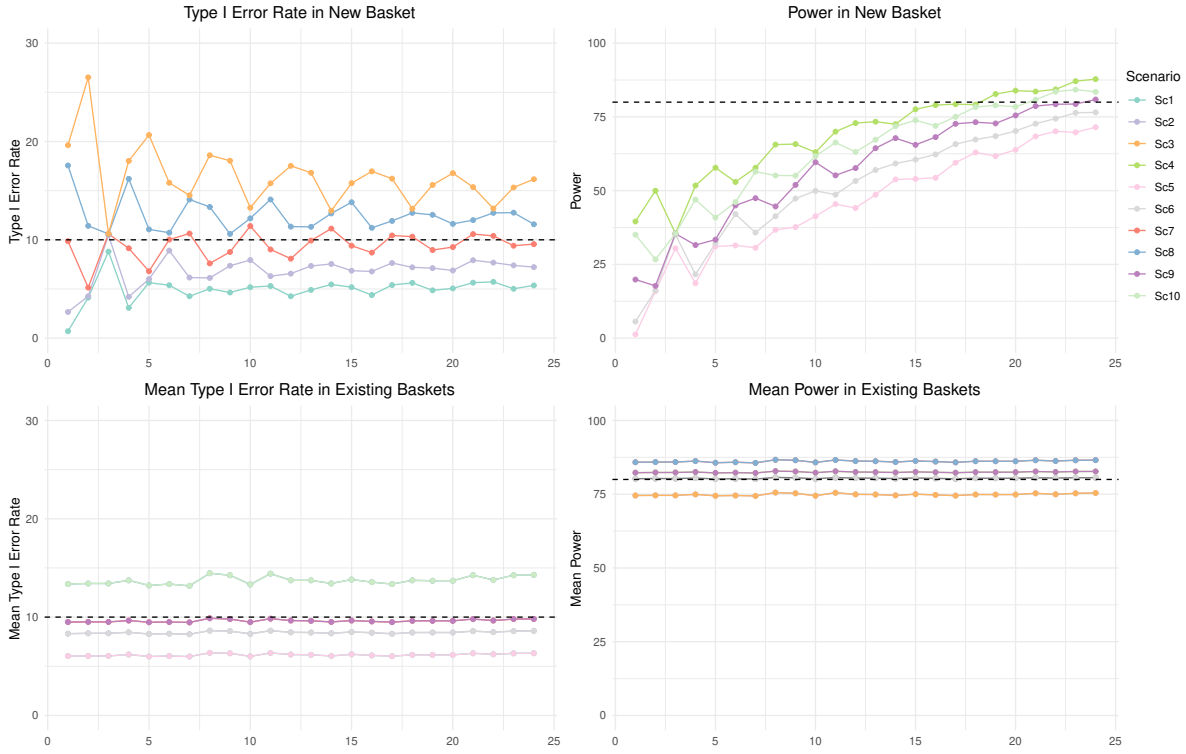


Figure B.6.2: Type I error rate and power under each sample size of n_5 from 1 to 24 by applying PL2(b), split by existing and new baskets.

To do so, again consider the fixed data scenario simulations setting with four existing and 1 new basket calibrated using the RCaP. In the previous simulation study, sample size of the new basket is assumed as known, consisting of $n_5 = 14$ patients, whilst existing baskets had $n_{k_0} = 24$ patients in each. Now assume n_5 is unknown.

First consider PL1(b) applied to all possible sample sizes from $n_5 = 1$ up to the full sample size of existing baskets, $n_5 = 24$, with separate calibrations of Δ_{k_0} and $\Delta_{k'}$ conducted for each value of n_5. Figure B.6.1 presents the type I error rate and power in new and existing baskets for each value of $n_5 = 1, \ldots, 24$ under scenarios 1-10 as presented in Table B.1.1.

Error rates and power for existing baskets are fairly consistent across all sample sizes, implying the time of addition of a new basket has little to no impact on the performance in these existing baskets, obviously an ideal characteristic.

However, much more noticeable changes are observed in the new basket. Cyclic fluctuations occur in both power and error rate due to the discreetness of data. As expected, power in the new basket generally increases with the sample size as more information is available. Across scenarios, power increases as the number of effective existing baskets also increases. The targeted 80% level is only reached under scenario 6 when $n_5 = 18$ and $n_5 = 21$ under scenario 7. The nominal level is never achieved under scenarios 7 and 8 in which there are none or just one effective existing basket.

In terms of error rates, more variation tends to occur when sample sizes are small, with the greatest error occurring when there are just 2 patients in the new basket (type I error rate of 23.8%) under scenario 5. As the number of effective existing baskets increases, the error rates are uniformly higher, with scenario 5 as the 'worst case' scenario where the only ineffective basket is the new basket. However, there is no general increase or decrease as the sample size increases and thus one cannot make the conclusion that any one sample size results in a more detrimental performance, at least in terms of the type I error rate.

The same interpretations are drawn when looking at the timing of addition under PL2(b) as plotted in Figure B.6.2. As PL1(b) and PL2(b) are equivalent for the new basket, with information borrowed between all baskets, the plots for type I error rate and power in the new basket are identical to that in Figure B.6.1. Again, little to no

variation is present in error and power for existing baskets as n_5 changes.

In summary, operating characteristics for PL1(b) and PL2(b) are fairly robust to the timing of addition of the new basket, particularly in the case of existing baskets, with little to no changes in power and error rate with the variation of sample sizes in the new basket. Power in the new basket is obviously improved as the sample size increases but no increase/decrease outside of the cyclic behaviour is observed in error rates, implying the type I error rate will be fairly unaffected by the timing of addition.

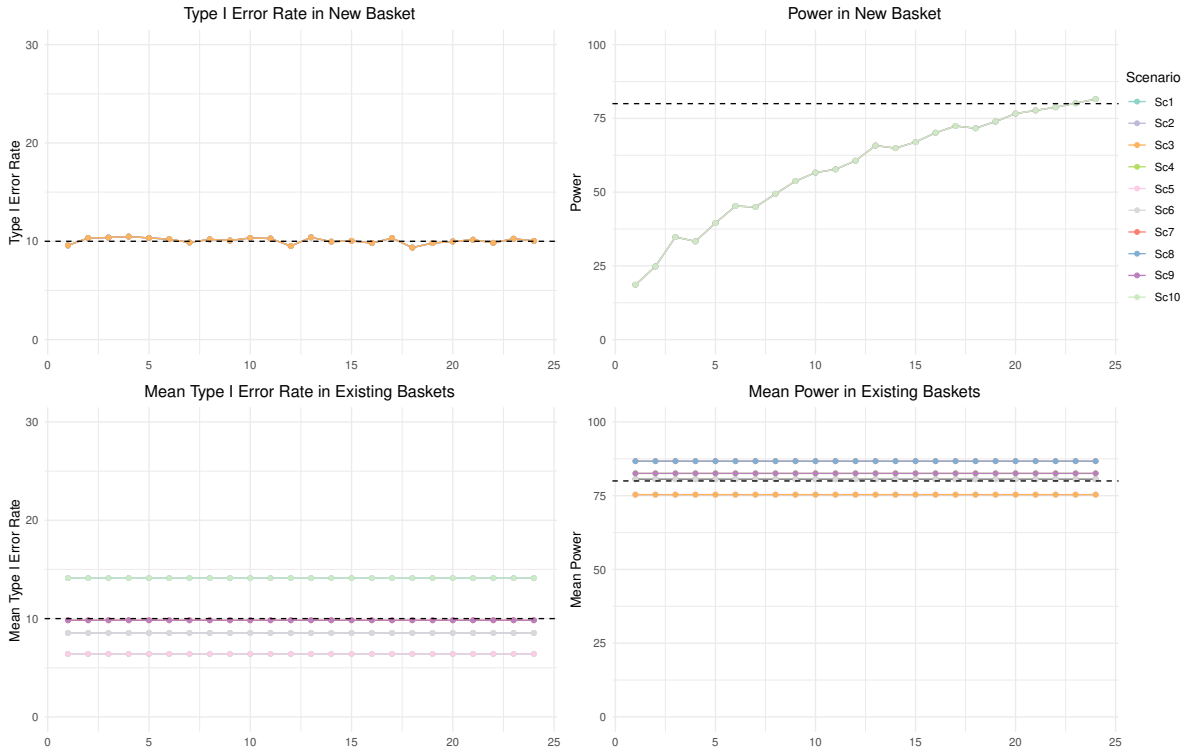


Figure B.6.3: Type I error rate and power under each sample size of n_5 from 1 to 24 by applying IND, split by existing and new baskets.

Under IND, timing of addition will not have an impact on existing baskets due to the independent analysis of new baskets. Thus, when sample sizes are smaller the only effect will be reduced power in the new basket with increased power as sample sizes grow. Under UNPL, the addition is not planned and so timing of addition has no relevance to the calibration procedure.

Figure B.6.3 presents the change in type I error rate and power as the sample size

in the new basket varies, split by new and existing baskets for an IND approach. As the new basket is analysed independently, the impact of its sample size on existing baskets is non-existent but also, as each sample size is calibrated to achieve 10% type I error rate, the impact of change in n_5 on error in the new basket is also null. The only variation is in power in the new basket, with larger sample sizes obviously improving power due to the increased certainty in posterior distributions from the added volume of information obtained.

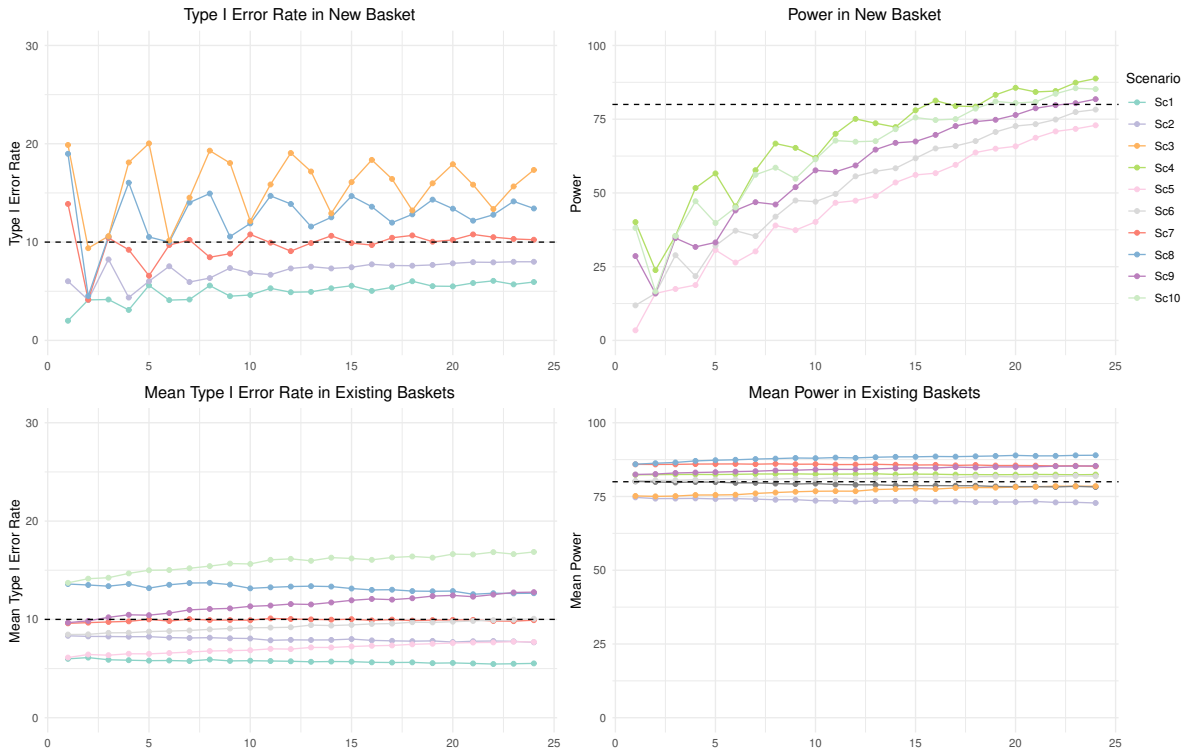


Figure B.6.4: Type I error rate and power under each sample size of n_5 from 1 to 24 by applying UNPL, split by existing and new baskets.

Now under UNPL, the new basket is an unplanned addition and thus the sample size of new baskets has no influence on the calibration procedure. Figure B.6.4 again presents change in type I error rate and power as n_5 varies. Results again imply the sample size in the new baskets has little to no impact on the performance in existing baskets with fairly consistent type I error rates and power across all n_5 values. Power in the new basket increases with the sample size as expected and type I error rates form

a cyclic pattern due to the discreetness of data. No general increase or decrease in type I error rate is observed as the sample size changes.

B.7 Simulation Study - 2 Existing Baskets with 2 New Baskets Added

All simulation studies conducted so far consisted of four existing baskets opening the trial with one additional basket added during the duration. Instead we now consider a case in which there are two baskets starting the trial with a further two baskets added at a later point.

Table B.7.1: Simulation study scenarios for the setting with 2 existing baskets with 2 new added.

	p_1	p_2	p_3	p_4
Scenario 1	0.2	0.2	0.2	0.2
Scenario 2	0.4	0.2	0.2	0.2
Scenario 3	0.4	0.4	0.2	0.2
Scenario 4	0.4	0.4	0.4	0.2
Scenario 5	0.2	0.2	0.4	0.2
Scenario 6	0.4	0.2	0.4	0.2
Scenario 7	0.2	0.2	0.4	0.4
Scenario 8	0.4	0.2	0.4	0.4
Scenario 9	0.4	0.4	0.4	0.4

The same design parameters as previously implemented are used here with a null and target response rate of $q_0 = 0.2$ and $q_1 = 0.4$ and a sample size of $n_{k_0} = 24$ in existing baskets and $n_{k\prime} = 14$ in newly added baskets. Models are specified as outlined in Section 3.5.2 and data scenarios considered are provided in Table B.7.1. Cut-off values Δ_{k_0} and $\Delta_{k\prime}$ under the IND, PL1(a) and PL2(a) approaches, are calibrated across scenarios 1-8 with Δ_{k_0} taken as the quantile of posterior probabilities across scenarios 1-2 and 5-8 for basket 2 and $\Delta_{k\prime}$ as the quantile across scenarios 1-6 of basket 4. For UNPL, the cut-off value is calibrated across just two scenarios: $p = (0.2, 0.2)$ and $p = (0.4, 0.2)$.

Now, as multiple baskets are added during the trial, the IND approach gives two options: (a) analyse both new baskets as independent of existing baskets and one another or (b) analyse both new baskets as independent of existing baskets but borrow from each other using a second EXNEX model.

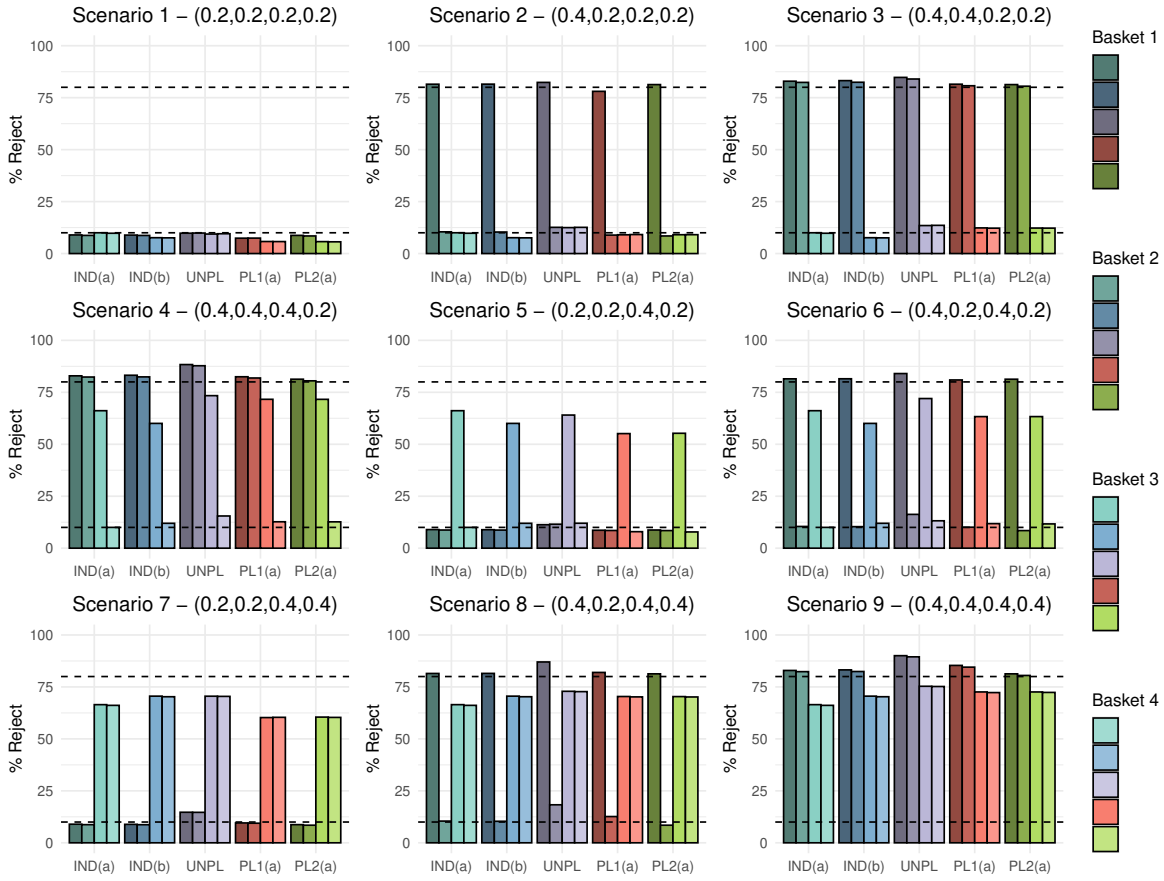


Figure B.7.1: Percentage of data sets within which the null hypothesis was rejected for a simulation study consisting of 2 existing baskets with 2 additional baskets added part-way through the study.

Results for the percentage of data sets within which the null hypothesis was rejected, i.e. type I error rate and power, are presented in Figure B.7.1. First, consider the differences between IND(a) and IND(b). Under IND(a), due to independent analysis, error rates in the new basket are always controlled to the 10% level but the same does not hold for IND(b) in which an EXNEX model allows borrowing between both new baskets. In cases where both new baskets are heterogeneous, this leads to reduced

error rates lying below the nominal level (e.g. 7.6% under scenario 1) but in cases of heterogeneity where one new basket is effective and the other ineffective, error rates inflate to approximately 12%. In these scenarios, power is pulled down from 66.2% to 60.1% when utilising information borrowing. However, significant power can be gained over an independent analysis in cases where both new baskets are effective to treatment (scenarios 7-9). Under IND(a) this power is 66.2% compared to IND(b) with power 70.6%.

Under UNPL, $\Delta_{k_0} = \Delta_{k'} = 0.865$ compared to 0.900 under PL1(a). This reduced cut-off value leads to less conservative rejections under both new and existing baskets. This results in higher power across all cases, with UNPL giving highest power in all scenarios (e.g. UNPL has a power of 89.8% and 75.3% for existing and new baskets respectively under data scenario 9, whereas, PL1(a) has power 84.9% and 72.4% respectively). With this, UNPL also possesses the greatest error inflation up to 18.3%.

Approaches PL1(a) and PL2(a) are equivalent for the new baskets so results differ only in existing baskets. Some cases with more noticeable differences are scenario 2 in which power is increased significantly under PL2(a) at 81.3% compared to 78.1% under PL1 with indistinguishable difference in error; scenario 8 in which both approaches give similar power but PL1 has higher error rates at 12.6% compared to PL2(a) at 8.5% and finally, scenario 9 in which power in existing baskets is greater under PL1(a) at 84.8% compared to 80.9% under PL2(a).

To conclude, results in this simulation study present fairly similar results to the previous case consisting of 4 existing and one new basket. One of the main differences lies in the UNPL approach which has a far less conservative cut-off than any of the other approaches. This occurs because there are only 2 existing baskets that can be used to calibrate UNPL and as such, estimates lack certainty and only 2 data scenarios are calibrated across. Also displayed in this case, is the potential losses one can make when utilising IND(b) in all cases bar when both baskets are homogeneous to treatment.

Appendix C

Supporting Information: Incorporating Historic Information to Further Improve Power When Conducting Bayesian Information Borrowing in Basket Trials

C.1 Robust Calibration Procedure (RCaP)

Algorithm 3 describes the Robust Calibration Procedure used within Chapter 4, specifically for the control of the type I error rate. Recall that in the simulation study in Chapter 4, calibration of efficacy thresholds, Δ_k, was conducted separately for the four historical data settings: (a) $y_{k^*} = (1, 1, 1)$, (b) $y_{k^*} = (3, 1, 1)$, (c) $y_{k^*} = (3, 3, 1)$ and (d) $y_{k^*} = (3, 3, 3)$. Therefore the RCaP was conducted four times with all 8 scenarios listed in Table 4.4.1 in Chapter 4 included. As sample sizes were equal, for baskets with identical historic data, Δ_k's are set as equal and to the basket whose RCaP calibrates

across the most scenarios. For example, in historic setting (b) in which the 2nd and 3rd baskets are identical, Δ_2 is set as the value of Δ_3 obtained from the RCaP procedure. This is because basket 3 is calibrated across scenarios 1, 2, 3, 7 and 8 (i.e. cases where the true response rate of p_3 is null), whereas, basket 3 is only null under scenarios 1, 2, 7 and 8, thus better error control is expected if Δ_3 is set as the efficacy cut-off for both baskets 2 and 3.

Algorithm 3 RCaP - Calibrate Δ_k across several simulation scenarios for type I error rate control

Data: Total number of simulation scenarios, M, scenarios $\boldsymbol{p_1}, \ldots, \boldsymbol{p_M}$, basket sample sizes $\boldsymbol{n_m}$, number of simulation runs for each scenario, R, null response rate, q_0 and integer weights for the scenarios, $\omega_1, \ldots, \omega_M$;
Initialisation: $\boldsymbol{Q_1}, \ldots, \boldsymbol{Q_K}$ empty vectors for storing Q
for $m = 1$ to M **do**
  for $r = 1$ to R **do**
    Generate data $\boldsymbol{X} \sim$ Binomial$(\boldsymbol{p_m}, \boldsymbol{n_m})$
    Fit information borrowing model to obtain posterior densities
    for $k = 1$ to K **do**
      Compute the posterior probability of a type I error $\mathbb{P}(p_{mk} > q_0|X)$, in basket k
      if $T(p_{mk} \leq q_0)$ **then**
        for $j = 1$ to ω_m **do**
          $\boldsymbol{Q_k} = \boldsymbol{Q_k} \cup \mathbb{P}(p_{mk} > q_0|X)$
        end for
      end if
    end for
  end for
end for
$\Delta_k = 100(1-\alpha)\%$ quantile of $\boldsymbol{Q}_k$ for each basket k.
return Cut-off values Δ_k for each basket k;

C.2 An Alternative Approach: EXNEX With SAM Prior in the NEX Component (EXsamNEX)

As described Chapter 4, selecting the power α in the power prior can be challenging and significantly effect inference. In the EXppNEX approach this power prior was placed on the NEX component in the EXNEX model. To avoid the specification of the power α, the power prior can be replaced with a self-adaptive mixture (SAM) prior:

$$
\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1, \ldots, K && \theta_k = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2),\\
&p_k = \delta_k M_{1k} + (1 - \delta_k) M_{2k} && \mu \sim \text{N}(m_\mu, \nu_\mu),\\
&\mathbb{I}_k = 1 \text{ if } y_{k^*}^{(j)} \text{ exists for basket } k \text{ for some } j \geq 1, && \sigma \sim g(\cdot),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && \tilde{\omega}_k \sim \text{Bernoulli}(\phi_k),\\
& && M_{2k} = \mathbb{I}_k \tilde{\omega}_k \pi_1(p_k) + (1 - \tilde{\omega}_k)\pi_0(p_k).
\end{aligned}
\tag{C.2.1}
$$

Note that the parameters for the EX component remain unchanged compared to the EXNEX and EXppNEX models. The SAM prior is placed on the NEX, M_{2k}, component consisting of a mixture of an informative prior, π_1, and uninformative prior, π_0. The non-informative prior is simply $\pi_0(p_k) = \text{Beta}(a_k, b_k)$ with values $a_k = b_k = 1$ recommended. The informative prior is a Meta-analytic predictive prior (Weber et al., 2021). The MAP prior in not tractable and thus MCMC methods would need to be utilised, however, it is approximated by a mixture of conjugate priors (Schmidli et al., 2014):

$$\pi_1(p_k) = \sum_{i=1}^{H_k} \kappa_i \text{Beta}(a_k + y_{k^*}^{(i)}, b_k + n_{k^*}^{(i)} - y_{k^*}^{(i)}),$$

where the κ_i weights are positive and sum to one. Weights can be defined as fixed in the model or can be updated in the posterior. Should there be a single source of historic data, this weight is set at $\kappa_k = 1$ thus $\pi_1(p_k) = \text{Beta}(a_k + y_{k^*}, b_k + n_{k^*} - y_{k^*})$.

The SAM prior mixture weights follow a Bernoulli distribution with probability ϕ_k, where ϕ_k is computed as guided by Yang et al. (2023), utilising the likelihood ratio test statistic. Let $\hat{p}_{k^*} = \int p_k \pi_1(p_k) dp_k$ be the expected value of p_k based on π_1. In cases of a single source of historic data in a basket $\hat{p}_{k^*} = (a_k + y_{k^*})/(a_k + b_k + n_{k^*})$ is the estimate of p_{k^*}, so the likelihood ratio test statistic is then:

$$R_k = \frac{\hat{p}_{k^*}^{y_k}(1-\hat{p}_{k^*})^{n_k - y_k}}{\max\{(\hat{p}_{k^*}+\Omega)^{y_k}(1-\hat{p}_{k^*}-\Omega)^{n_k-y_k}, (\hat{p}_{k^*}-\Omega)^{y_k}(1-\hat{p}_{k^*}+\Omega)^{n_k-y_k}\}}, \quad \text{(C.2.2)}$$

where ϕ_k is then set as $\phi_k = R_k/(1+R_k)$.

C.2.1 Simulation Study Model Specification

Within the simulation results presented in this Appendix, results of the EXsamNEX are also presented. The simulation setting is the same as that in Chapter 4, with 5 current baskets and historic information available for the first 3. The model applied is:

$$\begin{aligned}
&Y_k \sim \text{Binomial}(n_k, p_k), \qquad k = 1,2,3,4,5, && \theta_k = \text{logit}(M_{1k}) \sim \text{N}(\mu, \sigma^2),\\
&p_k = \delta_k M_{1k} + (1-\delta_k) M_{2k}, && \mu \sim \text{N}(\text{logit}(0.1), 10^2),\\
&\mathbb{I}_k = 1 \text{ if } y_{k^*} \text{ exists for basket } k, && \sigma \sim \text{Half-Normal}(0,1),\\
&\delta_k \sim \text{Bernoulli}(\pi_k), && \tilde{\omega}_k \sim \text{Bernoulli}(\phi_k),\\
& && M_{2k} = \mathbb{I}_k \tilde{\omega}_k \pi_1(p_k) + (1-\tilde{\omega}_k)\pi_0(p_k),\\
& && \pi_1(p_k) = \text{Beta}(1+y_{k^*}, 1+n_{k^*}-y_{k^*}),\\
& && \pi_0(p_k) = \text{Beta}(1,1),
\end{aligned}$$

where $\pi_k = 0.5$ for $k = 1,2,3,4,5$. Prior probabilities, ϕ_k are computed as follows: let $\hat{p}_{k^*} = (1+y_{k^*})/(1+1+n_{k^*})$ and

$$R_k = \frac{\hat{p}_{k^*}^{y_k}(1-\hat{p}_{k^*})^{n_k - y_k}}{\max\{(\hat{p_k^*}+0.15)^{y_k}(1-\hat{p}_{k^*}-0.15)^{n_k-y_k}, (\hat{p}_{k^*}-0.15)^{y_k}(1-\hat{p}_{k^*}+0.15)^{n_k-y_k}\}},$$

where 0.15 is the clinically relevant difference. $\phi_k = R_k/(1 + R_k)$.

C.3 Computational Time of Proposed Approaches

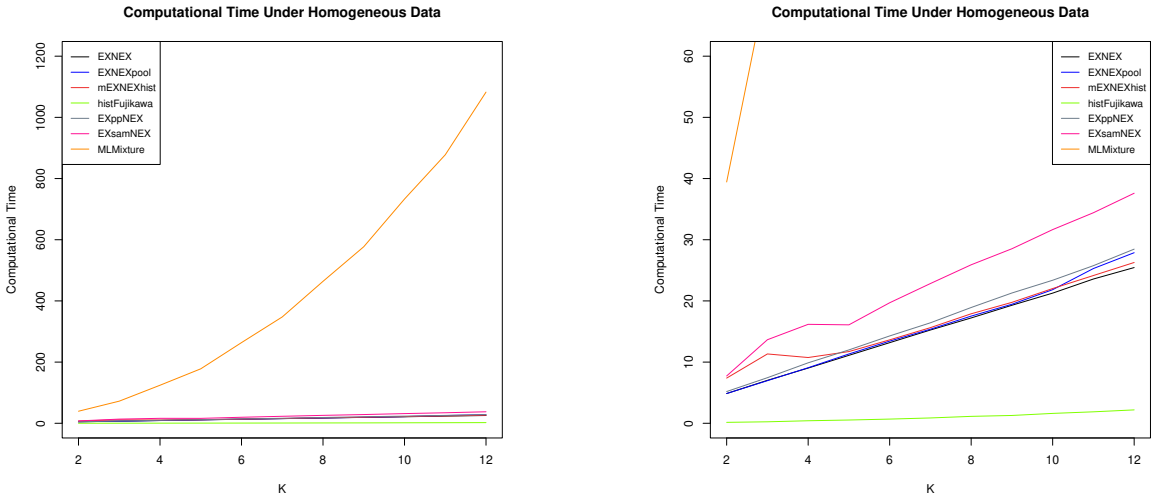
Each of the seven approaches explored in the simulation study vary in their model complexity and thus have varying computational intensity. For example, the Fujikawa$_{\text{hist}}$ approach has an analytical form, therefore does not require MCMC methods, making the model fit far quicker than all other approaches. In contrast the MLMixture model requires the mixture of two EXNEX models under which each of the K baskets are modelled separately, resulting in slow computation time. To add to this, the computation time will only increase as the total number of baskets on the trial increases (as demonstrated in Figures C.3.1 and C.3.2). The MLMixture models' computation time is further increased with the number of historic baskets also present.

Table C.3.1 presents the average computation time for each approach for a several fixed data sets. Each average is computed across 100 simulation runs for the same data with the standard deviation also presented. This is considered for five separate data sets, considering different combinations of effective/ineffective baskets and homogeneity/heterogeneity levels. From the results in Table C.3.1, it is clear that the data scenario has little effect on the computation time and thus it is expected that the only impacting factor will be the number of baskets present.

As expected, the histFujikawa approach takes a significantly shorter amount of time to conduct the model fit, averaging around half a second for each data set with a small standard deviation. At the opposite end of the spectrum, the MLMixture model takes around 15 times longer to fit the model compared to the standard EXNEX model. The EXNEX, EXNEX$_{\text{pool}}$, mEXNEX$_{\text{hist}}$ and EXppNEX approaches all take a similar amount of time, ranging from 10.8-12.1 seconds. The EXsamNEX model takes a couple of seconds longer at around 15.9 seconds due to the computation of mixture weights.

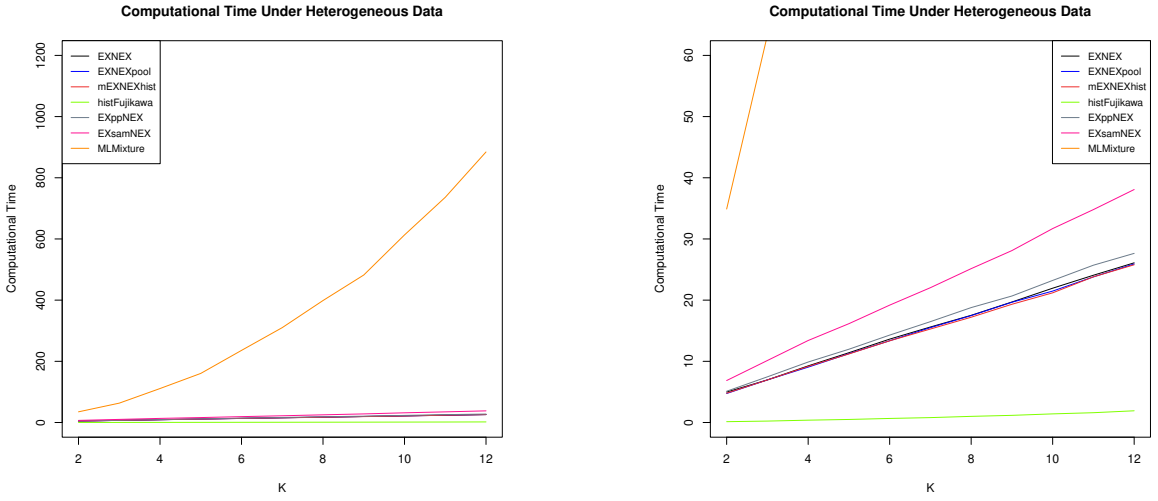
Table C.3.1: Computation time in seconds of all seven approaches for historic information borrowing measured in seconds. Each model is fit 100 times to the same data and the average computational time is taken and presented alongside the standard deviation. This is done for five different data sets (historic data available for the first three).

Method	$y_k = (3,3,3,3,3)$ $y_{k^*} = (1,1,1,0,0)$	$y_k = (9,9,3,3,3)$ $y_{k^*} = (3,3,1,0,0)$	$y_k = (9,9,3,3,3)$ $y_{k^*} = (1,1,3,0,0)$	$y_k = (9,9,9,9,9)$ $y_{k^*} = (3,3,3,0,0)$	$y_k = (9,9,9,9,9)$ $y_{k^*} = (3,1,1,0,0)$
EXNEX	11.556 (0.147)	11.279 (0.129)	11.012 (0.111)	11.024 (0.138)	11.105 (0.138)
$\text{EXNEX}_{\text{pool}}$	11.083 (0.132)	11.145 (0.101)	10.861 (0.107)	11.011 (0.126)	11.067 (0.155)
$\text{mEXNEX}_{\text{hist}}$	11.025 (0.128)	11.023 (0.110)	10.764 (0.101)	10.819 (0.103)	10.910 (0.140)
histFujikawa	0.538 (0.013)	0.468 (0.014)	0.481 (0.015)	0.444 (0.015)	0.471 (0.014)
EXppNEX	12.052 (0.126)	11.869 (0.151)	11.602 (0.137)	11.725 (0.160)	11.500 (0.129)
EXsamNEX	16.115 (0.160)	15.910 (0.147)	15.915 (0.111)	16.100 (0.157)	15.778 (0.110)
MLMixture	166.530 (1.380)	165.368 (0.641)	163.718 (1.315)	167.203 (1.793)	161.857 (0.684)



(a) (b)

Figure C.3.1: Average computational time of models fit on a fixed data set as K changes. Figure (b) is a zoomed-in version of (a) in order to distinguish the differences between methods. The fixed data set has all baskets homogeneous with current baskets each having a sample size of 34 with a total of 3 responses observed. Historic baskets have a sample size of 13 with 1 response observed. The number of historic baskets is $\lfloor K/2 \rfloor$.



(a) (b)

Figure C.3.2: Average computational time of models fit on a fixed data set as K changes. Figure (b) is a zoomed-in version of (a) in order to distinguish the differences between methods. The fixed data set has heterogeneity with even numbered baskets observing 9 responses and odd 3 responses. Historic baskets observe 1 response. Current baskets have a sample size of 34 and historic baskets have a sample size of 13. The number of historic baskets is $\lfloor K/2 \rfloor$.

C.4 Simulation Results

Table C.4.1: Simulation Results for Chapter 4 for scenario 1 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 1		**0.1**	**0.1**	**0.1**	**0.1**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		5.26	6.26	5.92	5.50	5.46	22.66	77.34
$\text{EXNEX}_{\text{pool}}$		6.10	6.20	6.42	5.44	5.40	23.12	76.88
$\text{mEXNEX}_{\text{hist}}$		3.84	4.00	3.26	2.74	2.68	11.44	88.56
histFujikawa		4.42	4.08	4.42	1.90	2.04	10.66	89.34
EXppNEX		7.24	6.94	6.94	.30	6.46	25.72	74.28
EXsamNEX		7.70	7.48	7.34	5.82	5.80	26.96	73.04
MLMixture		6.82	6.86	7.12	5.16	5.48	24.70	75.30
(b)	**(3,1,1)**							
EXNEX		5.26	6.26	5.92	5.50	5.46	22.66	77.34
$\text{EXNEX}_{\text{pool}}$		8.12	7.16	7.58	5.46	5.22	25.36	74.46
$\text{mEXNEX}_{\text{hist}}$		8.38	5.48	5.28	4.74	4.40	21.08	78.92
histFujikawa		7.02	4.42	4.08	2.16	2.32	12.64	87.36
EXppNEX		8.42	6.80	7.58	5.74	6.30	27.10	72.90
EXsamNEX		7.96	7.44	7.82	5.10	5.46	26.18	73.82
MLMixture		8.74	7.30	7.36	5.26	5.28	26.12	73.88
(c)	**(3,3,1)**							
EXNEX		5.26	6.26	5.92	5.50	5.46	22.66	77.34
$\text{EXNEX}_{\text{pool}}$		7.14	6.62	7.98	6.10	6.18	25.98	74.02
$\text{mEXNEX}_{\text{hist}}$		5.62	5.28	5.94	4.66	4.80	19.52	80.48
histFujikawa		4.50	4.96	4.66	2.84	2.86	13.00	87.00
EXppNEX		7.20	7.44	7.84	6.28	6.60	26.98	73.02
EXsamNEX		7.98	7.42	7.10	5.28	5.50	26.32	73.68
MLMixture		7.52	7.08	7.36	5.56	4.86	25.14	74.86
(d)	**(3,3,3)**							
EXNEX		5.26	6.26	5.92	5.50	5.46	22.66	77.34
$\text{EXNEX}_{\text{pool}}$		6.80	6.86	6.52	7.32	7.36	26.24	73.76
$\text{mEXNEX}_{\text{hist}}$		3.32	3.80	3.70	2.96	2.54	10.96	89.04
histFujikawa		4.94	4.66	4.68	3.04	2.62	13.82	86.18
EXppNEX		7.34	7.08	6.82	5.40	6.90	25.78	74.22
EXsamNEX		6.42	6.94	7.18	5.78	6.26	24.98	75.02
MLMixture		6.86	6.88	6.36	5.08	4.68	23.36	76.64

Table C.4.2: Simulation Results for Chapter 4 for scenario 2 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 2		**0.25**	**0.1**	**0.1**	**0.1**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		78.86	7.68	7.88	7.92	7.92	24.02	58.40
$\text{EXNEX}_{\text{pool}}$		80.30	9.28	9.48	6.36	6.86	24.36	58.72
$\text{mEXNEX}_{\text{hist}}$		66.92	8.84	9.12	6.30	5.80	21.8	47.04
histFujikawa		70.24	8.20	8.84	5.08	5.54	19.64	51.90
EXppNEX		81.24	9.52	9.68	8.32	7.96	27.78	56.92
EXsamNEX		83.00	10.32	9.48	6.88	7.02	26.96	58.82
MLMixture		81.12	9.42	9.52	6.02	6.68	25.42	58.88
(b)	**(3,1,1)**							
EXNEX		78.86	7.68	7.88	7.92	7.92	24.02	58.40
$\text{EXNEX}_{\text{pool}}$		81.44	8.90	8.50	6.86	6.66	23.70	60.72
$\text{mEXNEX}_{\text{hist}}$		82.96	8.02	8.32	6.70	6.80	22.18	62.74
histFujikawa		76.16	8.34	7.88	5.94	5.78	19.28	57.60
EXppNEX		83.68	9.66	9.44	8.10	7.70	27.46	59.36
EXsamNEX		82.76	9.24	8.82	7.40	6.74	25.96	59.54
MLMixture		83.06	9.30	9.04	6.90	6.56	25.40	60.74
(c)	**(3,3,1)**							
EXNEX		78.66	7.68	7.88	7.92	7.92	24.02	58.40
$\text{EXNEX}_{\text{pool}}$		81.94	10.16	9.38	7.52	8.22	27.18	57.88
$\text{mEXNEX}_{\text{hist}}$		77.62	10.82	9.28	8.46	7.44	27.50	52.90
histFujikawa		72.42	10.02	8.50	5.10	5.84	21.42	52.40
EXppNEX		81.76	10.28	9.10	8.10	7.56	28.28	56.40
EXsamNEX		82.08	9.18	10.26	6.70	6.46	26.32	58.80
MLMixture		82.02	9.40	8.72	6.90	6.92	24.86	60.04
(d)	**(3,3,3)**							
EXNEX		78.86	7.68	7.88	7.92	7.92	24.02	58.40
$\text{EXNEX}_{\text{pool}}$		81.06	8.98	9.64	9.18	8.92	28.94	56.14
$\text{mEXNEX}_{\text{hist}}$		68.06	9.38	8.48	5.80	6.48	21.60	48.44
histFujikawa		71.74	8.40	8.90	5.84	5.44	20.52	52.76
EXppNEX		81.70	9.68	9.48	8.88	7.68	28.18	56.78
EXsamNEX		80.84	9.22	8.82	6.74	6.80	24.74	59.52
MLMixture		81.10	8.60	9.26	6.54	7.10	24.96	59.54

Table C.4.3: Simulation Results for Chapter 4 for scenario 3 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 3		**0.25**	**0.25**	**0.1**	**0.1**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		82.54	83.16	10.70	10.76	11.26	28.66	45.40
$\text{EXNEX}_{\text{pool}}$		84.04	85.00	13.66	10.62	10.88	28.90	47.82
$\text{mEXNEX}_{\text{hist}}$		81.92	82.90	16.02	9.36	10.22	27.38	48.04
histFujikawa		81.24	81.04	13.88	9.76	10.68	26.12	45.66
EXppNEX		85.48	84.7	12.08	10.90	11.80	30.02	48.64
EXsamNEX		84.96	84.70	11.90	10.34	9.44	27.22	50.68
MLMixture		85.26	86.02	11.76	9.56	9.26	25.46	52.62
(b)	**(3,1,1)**							
EXNEX		82.54	83.16	10.70	10.76	11.26	28.66	45.40
$\text{EXNEX}_{\text{pool}}$		88.14	82.46	12.72	9.86	10.86	28.32	49.64
$\text{mEXNEX}_{\text{hist}}$		88.28	82.44	13.72	10.42	10.20	27.88	49.88
histFujikawa		86.02	81.58	14.32	11.18	10.28	27.32	47.96
EXppNEX		87.14	85.08	11.52	10.82	10.94	29.26	50.82
EXsamNEX		86.30	85.26	12.00	9.96	9.66	27.04	51.54
MLMixture		85.82	85.18	12.52	8.84	9.74	25.46	52.62
(c)	**(3,3,1)**							
EXNEX		82.54	83.16	10.70	10.76	11.26	28.66	45.40
$\text{EXNEX}_{\text{pool}}$		84.66	85.04	11.66	10.68	9.76	27.92	49.66
$\text{mEXNEX}_{\text{hist}}$		85.38	84.90	13.76	10.96	11.58	30.26	48.56
histFujikawa		82.74	82.70	13.02	10.14	9.74	25.34	48.08
EXppNEX		85.78	85.72	11.56	10.98	10.92	29.62	50.42
EXsamNEX		85.82	85.74	11.64	9.16	9.76	26.60	52.56
MLMixture		84.96	84.92	12.64	9.92	9.50	26.26	50.36
(d)	**(3,3,3)**							
EXNEX		82.54	83.16	10.70	10.76	11.26	28.66	45.40
$\text{EXNEX}_{\text{pool}}$		83.80	83.98	11.50	11.18	10.70	28.74	47.76
$\text{mEXNEX}_{\text{hist}}$		82.38	82.58	16.34	10.52	10.04	28.88	46.72
histFujikawa		82.56	82.12	13.16	10.00	10.30	25.46	47.66
EXppNEX		84.94	84.90	11.44	11.72	10.70	28.42	50.10
EXsamNEX		83.18	83.82	11.78	8.70	9.56	25.04	49.80
MLMixture		83.64	84.04	12.68	9.74	9.72	25.78	49.78

Table C.4.4: Simulation Results for Chapter 4 for scenario 4 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 4		**0.25**	**0.25**	**0.25**	**0.1**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		87.44	87.28	87.56	12.40	12.44	22.20	52.94
$\text{EXNEX}_{\text{pool}}$		88.64	89.38	88.92	14.12	13.34	23.56	53.98
$\text{mEXNEX}_{\text{hist}}$		89.52	89.26	89.52	15.38	14.86	25.26	52.86
histFujikawa		88.10	88.60	89.10	16.02	15.76	26.86	48.44
EXppNEX		88.02	87.36	88.08	11.26	11.62	21.36	54.08
EXsamNEX		88.34	88.18	88.28	12.36	12.28	21.44	54.38
MLMixture		88.64	88.42	88.68	13.30	13.34	22.86	53.66
(b)	**(3,1,1)**							
EXNEX		87.44	87.28	87.56	12.40	12.44	22.20	52.94
$\text{EXNEX}_{\text{pool}}$		90.90	89.30	88.84	13.72	13.58	23.52	55.04
$\text{mEXNEX}_{\text{hist}}$		92.30	89.76	89.08	13.28	13.94	23.70	56.26
histFujikawa		90.54	88.14	87.80	15.80	14.82	26.14	50.00
EXppNEX		89.46	88.26	88.40	11.84	12.60	23.00	53.90
EXsamNEX		88.30	87.84	88.18	12.78	12.92	22.24	53.14
MLMixture		89.46	88.56	88.56	12.98	13.22	21.80	54.08
(c)	**(3,3,1)**							
EXNEX		87.44	87.28	87.56	12.40	12.44	22.20	52.94
$\text{EXNEX}_{\text{pool}}$		90.02	88.74	86.64	11.32	12.36	21.58	54.00
$\text{mEXNEX}_{\text{hist}}$		90.52	89.58	89.16	12.84	13.08	22.92	54.74
histFujikawa		88.36	89.38	88.90	16.22	15.44	26.62	49.62
EXppNEX		88.28	87.50	88.70	11.60	12.00	21.46	53.94
EXsamNEX		88.58	88.24	88.94	13.14	12.68	22.34	53.68
MLMixture		88.78	89.60	87.98	13.48	12.02	22.08	53.96
(d)	**(3,3,3)**							
EXNEX		87.44	87.28	87.56	12.40	12.44	22.20	52.94
$\text{EXNEX}_{\text{pool}}$		88.34	89.24	87.86	10.84	11.74	21.20	55.28
$\text{mEXNEX}_{\text{hist}}$		89.20	90.00	89.46	15.74	14.92	25.86	51.82
histFujikawa		88.50	88.32	88.26	15.58	15.68	26.62	48.50
EXppNEX		88.82	87.62	87.70	12.22	12.26	21.14	54.02
EXsamNEX		88.80	87.62	87.70	12.22	12.26	21.14	54.02
MLMixture		87.86	87.38	88.84	12.86	13.04	21.92	53.14

Table C.4.5: Simulation Results for Chapter 4 for scenario 5 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 5		**0.25**	**0.25**	**0.25**	**0.25**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		88.48	89.48	88.86	88.64	14.48	14.48	52.34
$\text{EXNEX}_{\text{pool}}$		89.58	89.60	91.24	89.24	17.64	17.64	53.60
$\text{mEXNEX}_{\text{hist}}$		94.10	93.40	93.50	90.16	22.08	22.08	56.82
histFujikawa		91.68	91.68	92.16	88.16	20.46	20.46	54.18
EXppNEX		88.90	88.94	89.40	88.48	12.42	12.42	54.14
EXsamNEX		88.24	88.82	88.66	88.28	17.40	17.40	49.02
MLMixture		90.20	89.72	89.72	89.16	18.34	18.34	50.40
(b)	**(3,1,1)**							
EXNEX		88.48	89.48	88.86	88.64	14.48	14.48	52.34
$\text{EXNEX}_{\text{pool}}$		94.32	90.18	90.46	89.24	15.84	15.84	56.88
$\text{mEXNEX}_{\text{hist}}$		94.42	92.36	93.24	89.60	17.04	17.04	60.56
histFujikawa		93.06	91.96	91.60	88.26	20.16	20.16	54.72
EXppNEX		91.74	89.08	88.96	89.44	13.04	13.04	56.86
EXsamNEX		90.96	87.96	88.90	88.54	17.22	17.22	50.30
MLMixture		92.44	89.76	89.62	88.92	18.58	18.58	51.68
(c)	**(3,3,1)**							
EXNEX		88.48	89.48	88.86	88.64	14.48	14.48	52.34
$\text{EXNEX}_{\text{pool}}$		92.16	92.15	88.42	88.78	14.10	14.10	57.60
$\text{mEXNEX}_{\text{hist}}$		93.00	92.78	90.68	89.48	16.28	16.28	58.14
histFujikawa		92.30	92.40	91.300	89.78	19.10	19.10	57.00
EXppNEX		89.54	89.52	89.20	88.50	13.70	13.70	54.00
EXsamNEX		90.36	90.84	88.90	88.50	17.94	17.94	50.80
MLMixture		91.88	91.94	89.82	89.18	19.78	19.78	51.94
(d)	**(3,3,3)**							
EXNEX		88.48	89.48	88.86	88.64	14.48	14.48	52.34
$\text{EXNEX}_{\text{pool}}$		89.70	89.74	89.96	89.18	11.84	11.84	56.20
$\text{mEXNEX}_{\text{hist}}$		94.12	93.82	94.20	90.64	22.20	22.20	57.86
histFujikawa		91.68	92.08	90.96	88.74	19.66	19.66	54.66
EXppNEX		89.94	90.90	90.32	88.46	14.10	14.10	54.76
EXsamNEX		90.56	89.48	90.56	88.26	17.54	17.54	51.38
MLMixture		91.09	90.02	89.78	89.16	18.16	18.16	51.60

Table C.4.6: Simulation Results for Chapter 4 for scenario 6 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 6		**0.25**	**0.25**	**0.25**	**0.25**	**0.25**		
(a)	**(1,1,1)**							
EXNEX		90.16	89.82	90.06	90.28	90.10		61.10
$\text{EXNEX}_{\text{pool}}$		93.52	93.34	92.32	91.16	91.38		69.50
$\text{mEXNEX}_{\text{hist}}$		96.54	96.92	96.58	93.46	93.40		72.42
histFujikawa		94.24	94.54	94.78	92.62	92.84		75.60
EXppNEX		87.56	88.54	88.62	88.64	89.10		54.56
EXsamNEX		89.70	88.42	88.30	91.40	91.72		60.94
MLMixture		92.40	92.36	92.28	91.70	91.88		69.98
(b)	**(3,1,1)**							
EXNEX		90.16	89.82	90.06	90.28	90.10		61.10
$\text{EXNEX}_{\text{pool}}$		94.82	91.04	90.74	90.04	90.60		65.34
$\text{mEXNEX}_{\text{hist}}$		95.28	94.68	95.46	91.90	9.146		73.60
histFujikawa		95.14	94.22	94.44	92.18	92.64		75.38
EXppNEX		94.88	89.36	88.58	89.04	88.44		59.60
EXsamNEX		94.16	89.70	88.28	90.96	91.88		64.98
MLMixture		94.96	92.66	92.68	92.58	92.48		72.62
(c)	**(3,3,1)**							
EXNEX		90.16	89.82	90.06	90.28	90.10		61.10
$\text{EXNEX}_{\text{pool}}$		94.44	94.62	89.18	90.26	89.18		65.20
$\text{mEXNEX}_{\text{hist}}$		94.92	94.30	93.14	90.26	90.44		69.86
histFujikawa		94.94	94.70	94.38	93.00	91.96		75.90
EXppNEX		93.86	93.96	88.76	89.58	89.28		64.82
EXsamNEX		94.44	93.86	88.90	91.86	91.48		68.70
MLMixture		94.28	94.86	91.80	92.48	91.66		72.80
(d)	**(3,3,3)**							
EXNEX		90.16	89.82	90.06	90.28	90.10		61.10
$\text{EXNEX}_{\text{pool}}$		92.44	92.10	92.52	87.88	88.46		62.86
$\text{mEXNEX}_{\text{hist}}$		96.52	96.58	96.98	96.36	93.84		81.90
histFujikawa		93.70	94.56	94.70	92.80	92.26		75.42
EXppNEX		93.52	93.90	93.74	89.72	89.48		69.52
EXsamNEX		93.38	93.80	93.14	91.62	90.90		71.68
MLMixture		94.30	94.50	94.38	91.52	92.00		73.84

Table C.4.7: Simulation Results for Chapter 4 for scenario 7 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 6		**0.1**	**0.1**	**0.1**	**0.25**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		7.08	8.10	7.32	79.94	8.36	24.00	59.56
$\text{EXNEX}_{\text{pool}}$		8.04	7.94	8.48	78.36	6.12	23.54	58.66
$\text{mEXNEX}_{\text{hist}}$		7.18	7.04	7.66	63.84	5.42	18.90	48.12
histFujikawa		9.18	9.44	9.28	58.74	5.46	22.28	39.98
EXppNEX		10.40	10.58	10.70	80.00	8.76	32.58	51.68
EXsamNEX		10.84	10.76	10.24	77.74	7.40	32.12	51.40
MLMixture		9.58	9.80	9.34	79.20	6.88	28.06	55.80
(b)	**(3,1,1)**							
EXNEX		7.08	8.10	7.32	79.94	8.36	24.00	59.56
$\text{EXNEX}_{\text{pool}}$		11.88	9.64	9.46	79.18	7.48	29.92	53.88
$\text{mEXNEX}_{\text{hist}}$		11.62	9.14	9.74	78.90	7.24	27.24	53.00
histFujikawa		12.90	9.54	9.34	61.40	6.24	25.56	40.42
EXppNEX		11.58	10.24	9.92	79.32	8.12	32.80	51.42
EXsamNEX		12.04	10.76	9.54	79.64	7.54	32.04	52.34
MLMixture		11.26	9.28	9.36	77.82	7.50	29.18	53.48
(c)	**(3,3,1)**							
EXNEX		7.08	8.10	7.32	79.94	8.36	24.00	59.56
$\text{EXNEX}_{\text{pool}}$		10.08	10.18	8.94	79.40	8.28	29.54	53.58
$\text{mEXNEX}_{\text{hist}}$		8.46	8.66	8.58	75.92	7.16	24.36	54.94
histFujikawa		9.70	9.80	9.36	65.42	5.96	23.86	45.02
EXppNEX		10.22	9.88	9.74	79.88	8.62	30.90	53.08
EXsamNEX		10.70	10.76	9.60	79.60	7.12	30.24	54.04
MLMixture		10.58	10.30	9.62	77.08	7.00	28.74	53.06
(d)	**(3,3,3)**							
EXNEX		7.08	8.10	7.32	79.94	8.36	24.00	59.56
$\text{EXNEX}_{\text{pool}}$		10.10	9.40	9.92	81.86	8.70	30.08	55.10
$\text{mEXNEX}_{\text{hist}}$		71.80	7.18	7.32	65.44	4.78	18.42	49.54
histFujikawa		8.68	8.52	9.50	64.52	5.68	23.16	44.82
EXppNEX		9.70	10.22	10.00	80.36	8.52	30.90	53.78
EXsamNEX		9.70	9.44	9.14	80.08	7.34	27.80	56.70
MLMixture		9.26	10.06	9.08	77.36	7.36	27.10	54.34

Table C.4.8: Simulation Results for Chapter 4 for scenario 8 under historic cases (a), (b), (c) and (d).

	y_{k^*}	% Reject					FWER	% All Correct
Scenario 8		**0.25**	**0.1**	**0.1**	**0.25**	**0.1**		
(a)	**(1,1,1)**							
EXNEX		82.80	12.32	11.44	82.42	10.08	29.02	45.22
$\text{EXNEX}_{\text{pool}}$		83.06	12.02	11.96	81.24	9.76	27.74	45.14
$\text{mEXNEX}_{\text{hist}}$		78.66	14.28	13.94	76.96	8.94	27.50	41.56
histFujikawa		82.22	14.18	13.52	76.32	10.02	27.98	42.86
EXppNEX		86.72	11.72	10.60	82.74	10.96	29.54	49.42
EXsamNEX		86.16	11.72	11.04	82.22	10.66	30.06	47.40
MLMixture		86.10	11.92	12.26	82.12	10.02	28.72	48.42
(b)	**(3,1,1)**							
EXNEX		82.80	12.32	11.44	82.42	10.08	29.02	45.22
$\text{EXNEX}_{\text{pool}}$		88.20	12.22	11.72	81.10	10.36	29.22	48.08
$\text{mEXNEX}_{\text{hist}}$		86.68	13.04	13.78	80.68	10.10	28.14	47.42
histFujikawa		84.96	13.98	14.34	77.36	10.36	28.40	44.32
EXppNEX		88.50	11.76	11.54	83.24	11.30	30.74	49.72
EXsamNEX		86.74	11.96	11.80	81.30	10.46	30.03	47.92
MLMixture		87.16	12.26	11.72	80.06	9.12	27.48	48.56
(c)	**(3,3,1)**							
EXNEX		82.80	12.32	11.44	82.42	10.08	29.02	45.22
$\text{EXNEX}_{\text{pool}}$		86.94	13.04	12.04	81.82	11.10	30.18	47.26
$\text{mEXNEX}_{\text{hist}}$		82.98	15.24	12.44	81.48	9.66	30.12	44.22
histFujikawa		82.74	15.22	14.42	77.88	11.06	30.20	41.90
EXppNEX		86.18	12.40	11.76	83.52	10.60	30.22	48.52
EXsamNEX		86.24	12.62	11.40	81.40	10.54	29.44	47.64
MLMixture		85.90	13.22	11.64	80.80	9.88	28.34	47.40
(d)	**(3,3,3)**							
EXNEX		82.80	12.32	11.44	82.42	10.08	29.02	45.22
$\text{EXNEX}_{\text{pool}}$		84.66	12.12	12.42	83.94	10.74	29.94	47.96
$\text{mEXNEX}_{\text{hist}}$		79.12	14.30	14.16	77.58	9.22	27.90	41.82
histFujikawa		81.82	13.68	13.76	76.90	10.62	28.22	42.22
EXppNEX		86.14	11.88	12.26	82.06	9.72	28.96	48.24
EXsamNEX		83.62	12.72	13.08	81.82	10.22	29.56	45.66
MLMixture		84.74	12.48	12.62	80.44	9.94	28.22	45.82

Table C.4.9: Mean point estimate for the response rate (standard deviation) for scenario 1 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 1		**0.1**	**0.1**	**0.1**	**0.1**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.102 (0.039)	0.102 (0.041)	0.102 (0.040)	0.102 (0.039)	0.101 (0.040)
$\text{EXNEX}_{\text{pool}}$		0.096 (0.029)	0.096 (0.029)	0.095 (0.028)	0.101 (0.039)	0.101 (0.039)
$\text{mEXNEX}_{\text{hist}}$		0.100 (0.031)	0.100 (0.031)	0.100 (0.030)	0.101 (0.033)	0.101 (0.033)
histFujikawa		0.102 (0.029)	0.101 (0.030)	0.102 (0.030)	0.101 (0.029)	0.102 (0.030)
EXppNEX		0.107 (0.040)	0.108 (0.040)	0.107 (0.040)	0.101 (0.042)	0.101 (0.042)
EXsamNEX		0.104 (0.039)	0.104 (0.039)	0.105 (0.039)	0.105 (0.043)	0.105 (0.043)
MLMixture		0.103 (0.036)	0.104 (0.036)	0.104 (0.036)	0.103 (0.038)	0.103 (0.038)
(b)	**(3,1,1)**					
EXNEX		0.102 (0.039)	0.102 (0.041)	0.102 (0.040)	0.102 (0.039)	0.101 (0.040)
$\text{EXNEX}_{\text{pool}}$		0.129 (0.031)	0.099 (0.029)	0.099 (0.029)	0.103 (0.038)	0.102 (0.038)
$\text{mEXNEX}_{\text{hist}}$		0.102 (0.038)	0.101 (0.036)	0.101 (0.036)	0.102 (0.038)	0.102 (0.037)
histFujikawa		0.106 (0.033)	0.105 (0.032)	0.106 (0.032)	0.105 (0.032)	0.106 (0.033)
EXppNEX		0.113 (0.044)	0.107 (0.039)	0.107 (0.040)	0.100 (0.040)	0.102 (0.042)
EXsamNEX		0.107 (0.044)	0.104 (0.037)	0.104 (0.038)	0.104 (0.041)	0.104 (0.041)
MLMixture		0.113 (0.042)	0.105 (0.036)	0.105 (0.036)	0.105 (0.038)	0.105 (0.038)
(c)	**(3,3,1)**					
EXNEX		0.102 (0.039)	0.102 (0.041)	0.102 (0.040)	0.102 (0.039)	0.101 (0.040)
$\text{EXNEX}_{\text{pool}}$		0.130 (0.030)	0.130 (0.029)	0.101 (0.029)	0.106 (0.038)	0.105 (0.039)
$\text{mEXNEX}_{\text{hist}}$		0.101 (0.036)	0.100 (0.036)	0.101 (0.037)	0.101 (0.039)	0.101 (0.038)
histFujikawa		0.110 (0.036)	0.110 (0.036)	0.109 (0.037)	0.109 (0.036)	0.109 (0.036)
EXppNEX		0.113 (0.043)	0.113 (0.044)	0.107 (0.039)	0.100 (0.041)	0.101 (0.042)
EXsamNEX		0.107 (0.044)	0.107 (0.043)	0.104 (0.037)	0.103 (0.040)	0.104 (0.041)
MLMixture		0.114 (0.041)	0.114 (0.040)	0.107 (0.035)	0.107 (0.039)	0.107 (0.038)
(d)	**(3,3,3)**					
EXNEX		0.102 (0.039)	0.102 (0.041)	0.102 (0.040)	0.102 (0.039)	0.101 (0.040)
$\text{EXNEX}_{\text{pool}}$		0.134 (0.028)	0.133 (0.028)	0.133 (0.029)	0.109 (0.039)	0.109 (0.038)
$\text{mEXNEX}_{\text{hist}}$		0.100 (0.030)	0.100 (0.031)	0.100 (0.030)	0.100 (0.033)	0.100 (0.033)
histFujikawa		0.114 (0.040)	0.114 (0.039)	0.115 (0.040)	0.114 (0.040)	0.114 (0.039)
EXppNEX		0.113 (0.044)	0.113 (0.043)	0.112 (0.043)	0.099 (0.040)	0.100 (0.041)
EXsamNEX		0.106 (0.043)	0.107 (0.043)	0.107 (0.043)	0.104 (0.040)	0.105 (0.041)
MLMixture		0.115 (0.041)	0.116 (0.041)	0.115 (0.040)	0.109 (0.038)	0.109 (0.038)

Table C.4.10: Mean point estimate for the response rate (standard deviation) for scenario 2 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 2		**0.25**	**0.1**	**0.1**	**0.1**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.233 (0.074)	0.107 (0.043)	0.107 (0.043)	0.106 (0.043)	0.106 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.187 (0.052)	0.100 (0.031)	0.100 (0.031)	0.111 (0.040)	0.112 (0.040)
$\text{mEXNEX}_{\text{hist}}$		0.202 (0.066)	0.112 (0.036)	0.114 (0.036)	0.112 (0.038)	0.111 (0.037)
histFujikawa		0.214 (0.083)	0.106 (0.034)	0.107 (0.035)	0.106 (0.034)	0.106 (0.034)
EXppNEX		0.229 (0.064)	0.110 (0.041)	0.110 (0.041)	0.104 (0.044)	0.104 (0.043)
EXsamNEX		0.239 (0.077)	0.108 (0.040)	0.107 (0.040)	0.110 (0.044)	0.110 (0.045)
MLMixture		0.224 (0.071)	0.106 (0.038)	0.106 (0.037)	0.106 (0.039)	0.106 (0.040)
(b)	**(3,1,1)**					
EXNEX		0.233 (0.074)	0.107 (0.043)	0.107 (0.043)	0.106 (0.043)	0.106 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.229 (0.055)	0.100 (0.032)	0.100 (0.032)	0.106 (0.042)	0.105 (0.042)
$\text{mEXNEX}_{\text{hist}}$		0.230 (0.073)	0.106 (0.039)	0.106 (0.039)	0.106 (0.040)	0.105 (0.040)
histFujikawa		0.219 (0.076)	0.110 (0.037)	0.110 (0.037)	0.111 (0.037)	0.111 (0.037)
EXppNEX		0.246 (0.067)	0.110 (0.041)	0.109 (0.040)	0.104 (0.044)	0.104 (0.043)
EXsamNEX		0.243 (0.068)	0.107 (0.039)	0.106 (0.038)	0.109 (0.044)	0.108 (0.043)
MLMixture		0.239 (0.067)	0.108 (0.037)	0.107 (0.037)	0.109 (0.040)	0.110 (0.040)
(c)	**(3,3,1)**					
EXNEX		0.233 (0.074)	0.107 (0.043)	0.107 (0.043)	0.106 (0.043)	0.106 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.228 (0.053)	0.135 (0.032)	0.103 (0.032)	0.108 (0.042)	0.110 (0.043)
$\text{mEXNEX}_{\text{hist}}$		0.220 (0.070)	0.111 (0.040)	0.109 (0.042)	0.109 (0.041)	0.108 (0.041)
histFujikawa		0.222 (0.068)	0.116 (0.040)	0.116 (0.039)	0.114 (0.039)	0.115 (0.040)
EXppNEX		0.246 (0.066)	0.118 (0.045)	0.109 (0.041)	0.104 (0.044)	0.103 (0.043)
EXsamNEX		0.244 (0.069)	0.109 (0.045)	0.108 (0.039)	0.107 (0.043)	0.107 (0.042)
MLMixture		0.241 (0.067)	0.117 (0.041)	0.109 (0.037)	0.111 (0.040)	0.112 (0.040)
(d)	**(3,3,3)**					
EXNEX		0.233 (0.074)	0.107 (0.043)	0.107 (0.043)	0.106 (0.043)	0.106 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.226 (0.052)	0.139 (0.031)	0.139 (0.031)	0.112 (0.042)	0.112 (0.042)
$\text{mEXNEX}_{\text{hist}}$		0.202 (0.065)	0.114 (0.035)	0.113 (0.035)	0.111 (0.037)	0.111 (0.038)
histFujikawa		0.227 (0.063)	0.120 (0.043)	0.119 (0.044)	0.119 (0.043)	0.119 (0.042)
EXppNEX		0.245 (0.068)	0.117 (0.045)	0.116 (0.045)	0.104 (0.043)	0.102 (0.043)
EXsamNEX		0.242 (0.070)	0.110 (0.044)	0.111 (0.044)	0.107 (0.041)	0.108 (0.041)
MLMixture		0.239 (0.064)	0.120 (0.041)	0.121 (0.041)	0.114 (0.040)	0.115 (0.040)

Table C.4.11: Mean point estimate for the response rate (standard deviation) for scenario 3 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 3		**0.25**	**0.25**	**0.1**	**0.1**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.235 (0.069)	0.236 (0.069)	0.112 (0.046)	0.112 (0.046)	0.112 (0.046)
$\text{EXNEX}_{\text{pool}}$		0.190 (0.052)	0.191 (0.049)	0.105 (0.033)	0.111 (0.043)	0.112 (0.043)
$\text{mEXNEX}_{\text{hist}}$		0.218 (0.061)	0.218 (0.061)	0.126 (0.040)	0.120 (0.041)	0.120 (0.042)
histFujikawa		0.224 (0.075)	0.224 (0.075)	0.111 (0.039)	0.111 (0.038)	0.112 (0.039)
EXppNEX		0.232 (0.062)	0.230 (0.062)	0.114 (0.043)	0.108 (0.046)	0.109 (0.047)
EXsamNEX		0.239 (0.076)	0.238 (0.075)	0.113 (0.041)	0.116 (0.046)	0.116 (0.045)
MLMixture		0.228 (0.069)	0.228 (0.068)	0.110 (0.038)	0.113 (0.042)	0.112 (0.042)
(b)	**(3,1,1)**					
EXNEX		0.235 (0.069)	0.236 (0.069)	0.112 (0.046)	0.112 (0.046)	0.112 (0.046)
$\text{EXNEX}_{\text{pool}}$		0.230 (0.052)	0.193 (0.049)	0.105 (0.035)	0.112 (0.044)	0.112 (0.044)
$\text{mEXNEX}_{\text{hist}}$		0.230 (0.066)	0.226 (0.067)	0.118 (0.044)	0.114 (0.044)	0.114 (0.043)
histFujikawa		0.226 (0.067)	0.225 (0.068)	0.118 (0.041)	0.117 (0.041)	0.117 (0.040)
EXppNEX		0.247 (0.065)	0.230 (0.062)	0.113 (0.042)	0.108 (0.047)	0.108 (0.047)
EXsamNEX		0.245 (0.067)	0.240 (0.076)	0.110 (0.041)	0.114 (0.046)	0.114 (0.045)
MLMixture		0.239 (0.065)	0.228 (0.068)	0.112 (0.039)	0.114 (0.041)	0.116 (0.042)
(c)	**(3,3,1)**					
EXNEX		0.235 (0.069)	0.236 (0.069)	0.112 (0.046)	0.112 (0.046)	0.112 (0.046)
$\text{EXNEX}_{\text{pool}}$		0.232 (0.052)	0.233 (0.051)	0.105 (0.036)	0.113 (0.046)	0.113 (0.045)
$\text{mEXNEX}_{\text{hist}}$		0.228 (0.065)	0.228 (0.065)	0.116 (0.062)	0.116 (0.045)	0.117 (0.045)
histFujikawa		0.231 (0.064)	0.231 (0.064)	0.120 (0.044)	0.121 (0.044)	0.120 (0.044)
EXppNEX		0.248 (0.065)	0.248 (0.065)	0.112 (0.042)	0.107 (0.046)	0.107 (0.045)
EXsamNEX		0.244 (0.067)	0.246 (0.068)	0.109 (0.041)	0.112 (0.045)	0.113 (0.045)
MLMixture		0.240 (0.063)	0.241 (0.064)	0.115 (0.039)	0.117 (0.042)	0.117 (0.042)
(d)	**(3,3,3)**					
EXNEX		0.235 (0.069)	0.236 (0.069)	0.112 (0.046)	0.112 (0.046)	0.112 (0.046)
$\text{EXNEX}_{\text{pool}}$		0.231 (0.051)	0.230 (0.050)	0.144 (0.032)	0.116 (0.046)	0.115 (0.045)
$\text{mEXNEX}_{\text{hist}}$		0.217 (0.061)	0.219 (0.061)	0.126 (0.039)	0.121 (0.042)	0.121 (0.041)
histFujikawa		0.235 (0.059)	0.234 (0.058)	0.126 (0.048)	0.126 (0.048)	0.126 (0.048)
EXppNEX		0.246 (0.065)	0.247 (0.065)	0.120 (0.045)	0.107 (0.045)	0.106 (0.045)
EXsamNEX		0.243 (0.068)	0.243 (0.068)	0.115 (0.047)	0.111 (0.043)	0.113 (0.044)
MLMixture		0.241 (0.064)	0.241 (0.063)	0.125 (0.043)	0.121 (0.042)	0.119 (0.043)

Table C.4.12: Mean point estimate for the response rate (standard deviation) for scenario 4 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 4		**0.25**	**0.25**	**0.25**	**0.1**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.241 (0.067)	0.240 (0.066)	0.240 (0.065)	0.118 (0.047)	0.117 (0.049)
$\text{EXNEX}_{\text{pool}}$		0.194 (0.046)	0.195 (0.046)	0.195 (0.046)	0.118 (0.046)	0.117 (0.046)
$\text{mEXNEX}_{\text{hist}}$		0.230 (0.056)	0.229 (0.055)	0.230 (0.056)	0.131 (0.041)	0.131 (0.046)
histFujikawa		0.231 (0.068)	0.232 (0.068)	0.233 (0.068)	0.119 (0.043)	0.118 (0.042)
EXppNEX		0.234 (0.061)	0.233 (0.060)	0.233 (0.061)	0.112 (0.048)	0.113 (0.048)
EXsamNEX		0.240 (0.071)	0.241 (0.071)	0.240 (0.071)	0.121 (0.047)	0.121 (0.047)
MLMixture		0.229 (0.065)	0.229 (0.065)	0.230 (0.066)	0.119 (0.043)	0.119 (0.044)
(b)	**(3,1,1)**					
EXNEX		0.241 (0.067)	0.240 (0.066)	0.240 (0.065)	0.118 (0.047)	0.117 (0.049)
$\text{EXNEX}_{\text{pool}}$		0.230 (0.048)	0.198 (0.046)	0.198 (0.045)	0.119 (0.047)	0.119 (0.047)
$\text{mEXNEX}_{\text{hist}}$		0.240 (0.063)	0.235 (0.061)	0.235 (0.062)	0.122 (0.049)	0.123 (0.049)
histFujikawa		0.233 (0.063)	0.234 (0.062)	0.234 (0.063)	0.123 (0.046)	0.123 (0.045)
EXppNEX		0.250 (0.063)	0.233 (0.060)	0.234 (0.060)	0.113 (0.049)	0.113 (0.049)
EXsamNEX		0.245 (0.064)	0.240 (0.073)	0.240 (0.072)	0.121 (0.048)	0.121 (0.048)
MLMixture		0.243 (0.062)	0.230 (0.065)	0.230 (0.065)	0.122 (0.044)	0.122 (0.044)
(c)	**(3,3,1)**					
EXNEX		0.241 (0.067)	0.240 (0.066)	0.240 (0.065)	0.118 (0.047)	0.117 (0.049)
$\text{EXNEX}_{\text{pool}}$		0.235 (0.048)	0.233 (0.048)	0.199 (0.046)	0.118 (0.048)	0.118 (0.048)
$\text{mEXNEX}_{\text{hist}}$		0.237 (0.061)	0.235 (0.061)	0.239 (0.062)	0.122 (0.048)	0.123 (0.049)
histFujikawa		0.235 (0.058)	0.235 (0.057)	0.236 (0.057)	0.130 (0.048)	0.129 (0.048)
EXppNEX		0.248 (0.063)	0.248 (0.063)	0.233 (0.061)	0.111 (0.048)	0.111 (0.049)
EXsamNEX		0.246 (0.065)	0.245 (0.065)	0.242 (0.073)	0.120 (0.047)	0.120 (0.046)
MLMixture		0.242 (0.060)	0.244 (0.062)	0.232 (0.065)	0.126 (0.044)	0.124 (0.043)
(d)	**(3,3,3)**					
EXNEX		0.241 (0.067)	0.240 (0.066)	0.240 (0.065)	0.118 (0.047)	0.117 (0.049)
$\text{EXNEX}_{\text{pool}}$		0.237 (0.047)	0.236 (0.046)	0.235 (0.046)	0.119 (0.049)	0.119 (0.050)
$\text{mEXNEX}_{\text{hist}}$		0.229 (0.057)	0.230 (0.056)	0.228 (0.056)	0.132 (0.046)	0.139 (0.046)
histFujikawa		0.239 (0.053)	0.239 (0.052)	0.239 (0.053)	0.135 (0.052)	0.135 (0.052)
EXppNEX		0.250 (0.060)	0.250 (0.063)	0.247 (0.062)	0.111 (0.047)	0.111 (0.048)
EXsamNEX		0.247 (0.065)	0.245 (0.065)	0.245 (0.064)	0.118 (0.047)	0.118 (0.048)
MLMixture		0.244 (0.061)	0.243 (0.061)	0.243 (0.060)	0.127 (0.045)	0.128 (0.045)

Table C.4.13: Mean point estimate for the response rate (standard deviation) for scenario 5 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 5		**0.25**	**0.25**	**0.25**	**0.25**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.245 (0.062)	0.246 (0.062)	0.246 (0.062)	0.245 (0.062)	0.112 (0.053)
$\text{EXNEX}_{\text{pool}}$		0.201 (0.044)	0.201 (0.043)	0.201 (0.044)	0.235 (0.061)	0.123 (0.048)
$\text{mEXNEX}_{\text{hist}}$		0.241 (0.052)	0.239 (0.051)	0.240 (0.051)	0.241 (0.053)	0.140 (0.052)
histFujikawa		0.237 (0.063)	0.237 (0.064)	0.238 (0.064)	0.235 (0.063)	0.126 (0.046)
EXppNEX		0.240 (0.059)	0.238 (0.058)	0.239 (0.058)	0.246 (0.064)	0.119 (0.052)
EXsamNEX		0.244 (0.068)	0.244 (0.067)	0.244 (0.068)	0.250 (0.065)	0.131 (0.050)
MLMixture		0.235 (0.064)	0.235 (0.064)	0.234 (0.063)	0.245 (0.067)	0.128 (0.046)
(b)	**(3,1,1)**					
EXNEX		0.245 (0.062)	0.246 (0.062)	0.246 (0.062)	0.245 (0.062)	0.112 (0.053)
$\text{EXNEX}_{\text{pool}}$		0.235 (0.045)	0.204 (0.044)	0.204 (0.043)	0.240 (0.061)	0.123 (0.050)
$\text{mEXNEX}_{\text{hist}}$		0.245 (0.059)	0.242 (0.056)	0.244 (0.056)	0.244 (0.058)	0.126 (0.052)
histFujikawa		0.237 (0.058)	0.239 (0.058)	0.239 (0.058)	0.238 (0.058)	0.131 (0.049)
EXppNEX		0.253 (0.058)	0.239 (0.059)	0.239 (0.049)	0.247 (0.065)	0.119 (0.053)
EXsamNEX		0.253 (0.057)	0.250 (0.064)	0.248 (0.065)	0.255 (0.059)	0.256 (0.059)
MLMixture		0.244 (0.059)	0.234 (0.063)	0.234 (0.062)	0.246 (0.066)	0.129 (0.047)
(c)	**(3,3,1)**					
EXNEX		0.245 (0.062)	0.246 (0.062)	0.246 (0.062)	0.245 (0.062)	0.112 (0.053)
$\text{EXNEX}_{\text{pool}}$		0.238 (0.045)	0.238 (0.044)	0.206 (0.044)	0.242 (0.060)	0.125 (0.052)
$\text{mEXNEX}_{\text{hist}}$		0.242 (0.056)	0.242 (0.057)	0.242 (0.059)	0.245 (0.058)	0.126 (0.052)
histFujikawa		0.241 (0.054)	0.242 (0.053)	0.241 (0.054)	0.242 (0.053)	0.137 (0.052)
EXppNEX		0.251 (0.060)	0.251 (0.060)	0.238 (0.058)	0.243 (0.066)	0.119 (0.051)
EXsamNEX		0.249 (0.062)	0.250 (0.061)	0.245 (0.070)	0.249 (0.065)	0.129 (0.049)
MLMixture		0.246 (0.057)	0.246 (0.057)	0.235 (0.060)	0.247 (0.064)	0.135 (0.047)
(d)	**(3,3,3)**					
EXNEX		0.245 (0.062)	0.246 (0.062)	0.246 (0.062)	0.245 (0.062)	0.112 (0.053)
$\text{EXNEX}_{\text{pool}}$		0.241 (0.044)	0.241 (0.043)	0.241 (0.044)	0.244 (0.058)	0.124 (0.053)
$\text{mEXNEX}_{\text{hist}}$		0.241 (0.052)	0.240 (0.051)	0.240 (0.051)	0.242 (0.054)	0.139 (0.052)
histFujikawa		0.244 (0.050)	0.245 (0.049)	0.244 (0.050)	0.244 (0.049)	0.145 (0.055)
EXppNEX		0.252 (0.060)	0.252 (0.059)	0.253 (0.060)	0.243 (0.067)	0.116 (0.051)
EXsamNEX		0.250 (0.060)	0.249 (0.063)	0.249 (0.061)	0.250 (0.068)	0.128 (0.050)
MLMixture		0.248 (0.056)	0.246 (0.057)	0.246 (0.056)	0.248 (0.062)	0.135 (0.049)

Table C.4.14: Mean point estimate for the response rate (standard deviation) for scenario 6 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 6		**0.25**	**0.25**	**0.25**	**0.25**	**0.25**
(a)	**(1,1,1)**					
EXNEX		0.251 (0.058)	0.250 (0.058)	0.250 (0.057)	0.251 (0.058)	0.250 (0.058)
$\text{EXNEX}_{\text{pool}}$		0.208 (0.042)	0.207 (0.042)	0.206 (0.043)	0.241 (0.058)	0.240 (0.057)
$\text{mEXNEX}_{\text{hist}}$		0.250 (0.046)	0.250 (0.045)	0.251 (0.046)	0.250 (0.049)	0.249 (0.048)
histFujikawa		0.241 (0.057)	0.242 (0.058)	0.242 (0.057)	0.242 (0.057)	0.243 (0.057)
EXppNEX		0.242 (0.057)	0.243 (0.056)	0.243 (0.056)	0.252 (0.062)	0.251 (0.060)
EXsamNEX		0.250 (0.064)	0.247 (0.065)	0.247 (0.064)	0.256 (0.060)	0.256 (0.060)
MLMixture		0.238 (0.060)	0.239 (0.06)	0.238 (0.060)	0.248 (0.060)	0.250 (0.062)
(b)	**(3,1,1)**					
EXNEX		0.251 (0.058)	0.250 (0.058)	0.250 (0.057)	0.251 (0.058)	0.250 (0.058)
$\text{EXNEX}_{\text{pool}}$		0.239 (0.042)	0.209 (0.043)	0.210 (0.043)	0.243 (0.057)	0.242 (0.057)
$\text{mEXNEX}_{\text{hist}}$		0.250 (0.055)	0.248 (0.051)	0.250 (0.051)	0.250 (0.054)	0.249 (0.054)
histFujikawa		0.244 (0.053)	0.245 (0.052)	0.245 (0.053)	0.243 (0.053)	0.245 (0.053)
EXppNEX		0.255 (0.055)	0.243 (0.055)	0.243 (0.056)	0.250 (0.061)	0.104 (0.061)
EXsamNEX		0.114 (0.047)	0.108 (0.040)	0.108 (0.039)	0.224 (0.076)	0.110 (0.044)
MLMixture		0.250 (0.055)	0.239 (0.057)	0.239 (0.057)	0.248 (0.058)	0.250 (0.059)
(c)	**(3,3,1)**					
EXNEX		0.251 (0.058)	0.250 (0.058)	0.250 (0.057)	0.251 (0.058)	0.250 (0.058)
$\text{EXNEX}_{\text{pool}}$		0.243 (0.042)	0.242 (0.042)	0.213 (0.043)	0.246 (0.054)	0.246 (0.056)
$\text{mEXNEX}_{\text{hist}}$		0.251 (0.052)	0.250 (0.053)	0.250 (0.055)	0.250 (0.055)	0.251 (0.056)
histFujikawa		0.248 (0.048)	0.247 (0.048)	0.248 (0.050)	0.248 (0.049)	0.248 (0.049)
EXppNEX		0.255 (0.056)	0.256 (0.057)	0.243 (0.056)	0.250 (0.062)	0.250 (0.061)
EXsamNEX		0.254 (0.057)	0.254 (0.057)	0.250 (0.067)	0.255 (0.060)	0.255 (0.061)
MLMixture		0.251 (0.054)	0.252 (0.054)	0.241 (0.057)	0.252 (0.058)	0.251 (0.059)
(d)	**(3,3,3)**					
EXNEX		0.251 (0.058)	0.250 (0.058)	0.250 (0.057)	0.251 (0.058)	0.250 (0.058)
$\text{EXNEX}_{\text{pool}}$		0.244 (0.041)	0.245 (0.041)	0.245 (0.040)	0.247 (0.056)	0.248 (0.055)
$\text{mEXNEX}_{\text{hist}}$		0.248 (0.045)	0.249 (0.046)	0.249 (0.046)	0.248 (0.048)	0.249 (0.048)
histFujikawa		0.250 (0.045)	0.251 (0.045)	0.251 (0.045)	0.250 (0.044)	0.250 (0.045)
EXppNEX		0.254 (0.056)	0.254 (0.057)	0.255 (0.056)	0.249 (0.062)	0.248 (0.061)
EXsamNEX		0.253 (0.058)	0.253 (0.057)	0.252 (0.058)	0.255 (0.062)	0.254 (0.061)
MLMixture		0.252 (0.053)	0.253 (0.052)	0.252 (0.053)	0.254 (0.055)	0.254 (0.057)

Table C.4.15: Mean point estimate for the response rate (standard deviation) for scenario 7 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 7		**0.1**	**0.1**	**0.1**	**0.25**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.105 (0.042)	0.106 (0.043)	0.106 (0.042)	0.233 (0.073)	0.107 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.098 (0.030)	0.099 (0.030)	0.098 (0.031)	0.232 (0.074)	0.104 (0.040)
$\text{mEXNEX}_{\text{hist}}$		0.109 (0.035)	0.109 (0.034)	0.110 (0.035)	0.213 (0.072)	0.108 (0.037)
histFujikawa		0.107 (0.034)	0.106 (0.034)	0.107 (0.034)	0.216 (0.084)	0.106 (0.034)
EXppNEX		0.111 (0.041)	0.112 (0.041)	0.112 (0.041)	0.235 (0.073)	0.105 (0.044)
EXsamNEX		0.108 (0.041)	0.109 (0.040)	0.107 (0.040)	0.242 (0.076)	0.110 (0.044)
MLMixture		0.107 (0.036)	0.106 (0.037)	0.107 (0.036)	0.240 (0.077)	0.107 (0.040)
(b)	**(3,1,1)**					
EXNEX		0.105 (0.042)	0.106 (0.043)	0.106 (0.042)	0.233 (0.073)	0.107 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.134 (0.032)	0.102 (0.031)	0.101 (0.031)	0.231 (0.073)	0.108 (0.041)
$\text{mEXNEX}_{\text{hist}}$		0.107 (0.040)	0.107 (0.039)	0.107 (0.039)	0.228 (0.074)	0.107 (0.041)
histFujikawa		0.112 (0.037)	0.111 (0.037)	0.112 (0.037)	0.219 (0.075)	0.111 (0.038)
EXppNEX		0.119 (0.045)	0.111 (0.042)	0.110 (0.041)	0.234 (0.074)	0.104 (0.043)
EXsamNEX		0.245 (0.066)	0.110 (0.042)	0.111 (0.040)	0.246 (0.073)	0.116 (0.047)
MLMixture		0.116 (0.041)	0.109 (0.037)	0.108 (0.037)	0.240 (0.078)	0.111 (0.040)
(c)	**(3,3,1)**					
EXNEX		0.105 (0.042)	0.106 (0.043)	0.106 (0.042)	0.233 (0.073)	0.107 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.135 (0.030)	0.136 (0.031)	0.103 (0.031)	0.230 (0.072)	0.109 (0.041)
$\text{mEXNEX}_{\text{hist}}$		0.106 (0.039)	0.107 (0.039)	0.106 (0.041)	0.230 (0.074)	0.107 (0.041)
histFujikawa		0.116 (0.040)	0.116 (0.040)	0.116 (0.041)	0.221 (0.068)	0.115 (0.040)
EXppNEX		0.117 (0.045)	0.117 (0.045)	0.110 (0.040)	0.233 (0.074)	0.105 (0.044)
EXsamNEX		0.111 (0.045)	0.112 (0.045)	0.108 (0.039)	0.242 (0.076)	0.109 (0.042)
MLMixture		0.118 (0.042)	0.118 (0.041)	0.110 (0.038)	0.238 (0.075)	0.112 (0.040)
(d)	**(3,3,3)**					
EXNEX		0.105 (0.042)	0.106 (0.043)	0.106 (0.042)	0.233 (0.073)	0.107 (0.043)
$\text{EXNEX}_{\text{pool}}$		0.139 (0.030)	0.139 (0.029)	0.139 (0.030)	0.229 (0.070)	0.114 (0.040)
$\text{mEXNEX}_{\text{hist}}$		0.109 (0.035)	0.109 (0.035)	0.109 (0.035)	0.214 (0.072)	0.108 (0.037)
histFujikawa		0.119 (0.044)	0.118 (0.043)	0.120 (0.043)	0.225 (0.061)	0.119 (0.043)
EXppNEX		0.116 (0.044)	0.117 (0.045)	0.118 (0.044)	0.234 (0.075)	0.104 (0.042)
EXsamNEX		0.111 (0.045)	0.111 (0.044)	0.111 (0.044)	0.243 (0.077)	0.108 (0.042)
MLMixture		0.121 (0.042)	0.121 (0.042)	0.120 (0.042)	0.238 (0.075)	0.114 (0.041)

Table C.4.16: Mean point estimate for the response rate (standard deviation) for scenario 8 under historic cases (a), (b), (c) and (d) for the simulation study in Chapter 4.

	y_{k^*}	Mean Point Estimate (Sd)				
Scenario 8		**0.25**	**0.1**	**0.1**	**0.25**	**0.1**
(a)	**(1,1,1)**					
EXNEX		0.237 (0.070)	0.113 (0.046)	0.112 (0.046)	0.235 (0.069)	0.112 (0.045)
$\text{EXNEX}_{\text{pool}}$		0.191 (0.051)	0.103 (0.033)	0.103 (0.033)	0.232 (0.070)	0.110 (0.043)
$\text{mEXNEX}_{\text{hist}}$		0.214 (0.062)	0.123 (0.039)	0.122 (0.039)	0.222 (0.066)	0.118 (0.040)
histFujikawa		0.226 (0.076)	0.111 (0.038)	0.112 (0.034)	0.226 (0.075)	0.112 (0.038)
EXppNEX		0.233 (0.061)	0.114 (0.043)	0.114 (0.042)	0.238 (0.072)	0.109 (0.046)
EXsamNEX		0.241 (0.076)	0.112 (0.042)	0.110 (0.040)	0.247 (0.072)	0.117 (0.047)
MLMixture		0.227 (0.068)	0.111 (0.038)	0.110 (0.039)	0.241 (0.074)	0.113 (0.042)
(b)	**(3,1,1)**					
EXNEX		0.237 (0.070)	0.113 (0.046)	0.112 (0.046)	0.235 (0.069)	0.112 (0.045)
$\text{EXNEX}_{\text{pool}}$		0.233 (0.052)	0.104 (0.034)	0.104 (0.034)	0.234 (0.070)	0.111 (0.044)
$\text{mEXNEX}_{\text{hist}}$		0.232 (0.070)	0.115 (0.043)	0.115 (0.043)	0.230 (0.069)	0.112 (0.044)
histFujikawa		0.226 (0.068)	0.116 (0.041)	0.117 (0.041)	0.229 (0.069)	0.116 (0.041)
EXppNEX		0.249 (0.065)	0.114 (0.043)	0.114 (0.042)	0.239 (0.071)	0.110 (0.047)
EXsamNEX		0.244 (0.067)	0.116 (0.047)	0.110 (0.040)	0.245 (0.074)	0.115 (0.046)
MLMixture		0.241 (0.065)	0.112 (0.039)	0.112 (0.039)	0.240 (0.075)	0.115 (0.041)
(c)	**(3,3,1)**					
EXNEX		0.237 (0.070)	0.113 (0.046)	0.112 (0.046)	0.235 (0.069)	0.112 (0.045)
$\text{EXNEX}_{\text{pool}}$		0.232 (0.051)	0.142 (0.033)	0.107 (0.035)	0.232 (0.068)	0.114 (0.045)
$\text{mEXNEX}_{\text{hist}}$		0.226 (0.068)	0.118 (0.044)	0.114 (0.044)	0.232 (0.069)	0.113 (0.044)
histFujikawa		0.230 (0.063)	0.123 (0.044)	0.122 (0.044)	0.230 (0.063)	0.123 (0.044)
EXppNEX		0.248 (0.065)	0.124 (0.046)	0.114 (0.042)	0.237 (0.071)	0.108 (0.046)
EXsamNEX		0.244 (0.045)	0.116 (0.047)	0.110 (0.040)	0.245 (0.074)	0.115 (0.046)
MLMixture		0.241 (0.063)	0.124 (0.043)	0.114 (0.039)	0.239 (0.072)	0.117 (0.042)
(d)	**(3,3,3)**					
EXNEX		0.237 (0.070)	0.113 (0.046)	0.112 (0.046)	0.235 (0.069)	0.112 (0.045)
$\text{EXNEX}_{\text{pool}}$		0.231 (0.051)	0.145 (0.032)	0.145 (0.032)	0.233 (0.067)	0.117 (0.045)
$\text{mEXNEX}_{\text{hist}}$		0.214 (0.062)	0.123 (0.039)	0.123 (0.039)	0.223 (0.065)	0.118 (0.040)
histFujikawa		0.233 (0.057)	0.126 (0.048)	0.127 (0.047)	0.233 (0.057)	0.126 (0.048)
EXppNEX		0.246 (0.064)	0.123 (0.046)	0.123 (0.045)	0.235 (0.072)	0.106 (0.044)
EXsamNEX		0.243 (0.068)	0.116 (0.046)	0.116 (0.047)	0.243 (0.074)	0.114 (0.044)
MLMixture		0.241 (0.063)	0.126 (0.043)	0.126 (0.043)	0.238 (0.070)	0.121 (0.043)

C.5 Exploring the Choice of Power, α, in the EXpp-NEX Approach

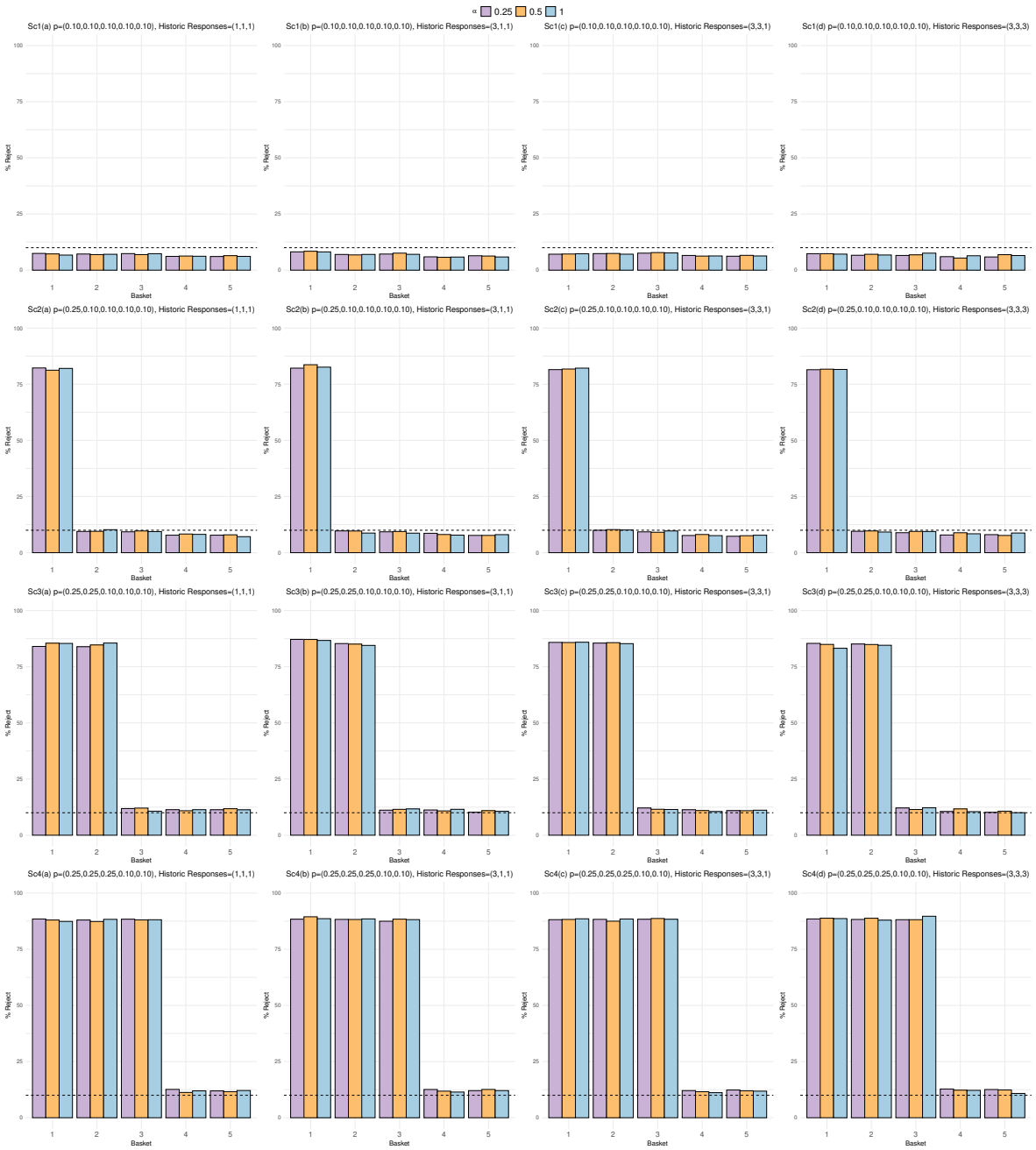


Figure C.5.1: The percentage of data sets where the null hypothesis were rejected per baskets under the EXppNEX model for scenarios 1-4 and 4 historic sub-cases. This is provided for three choices of α: 0.25, 0.5 and 1.

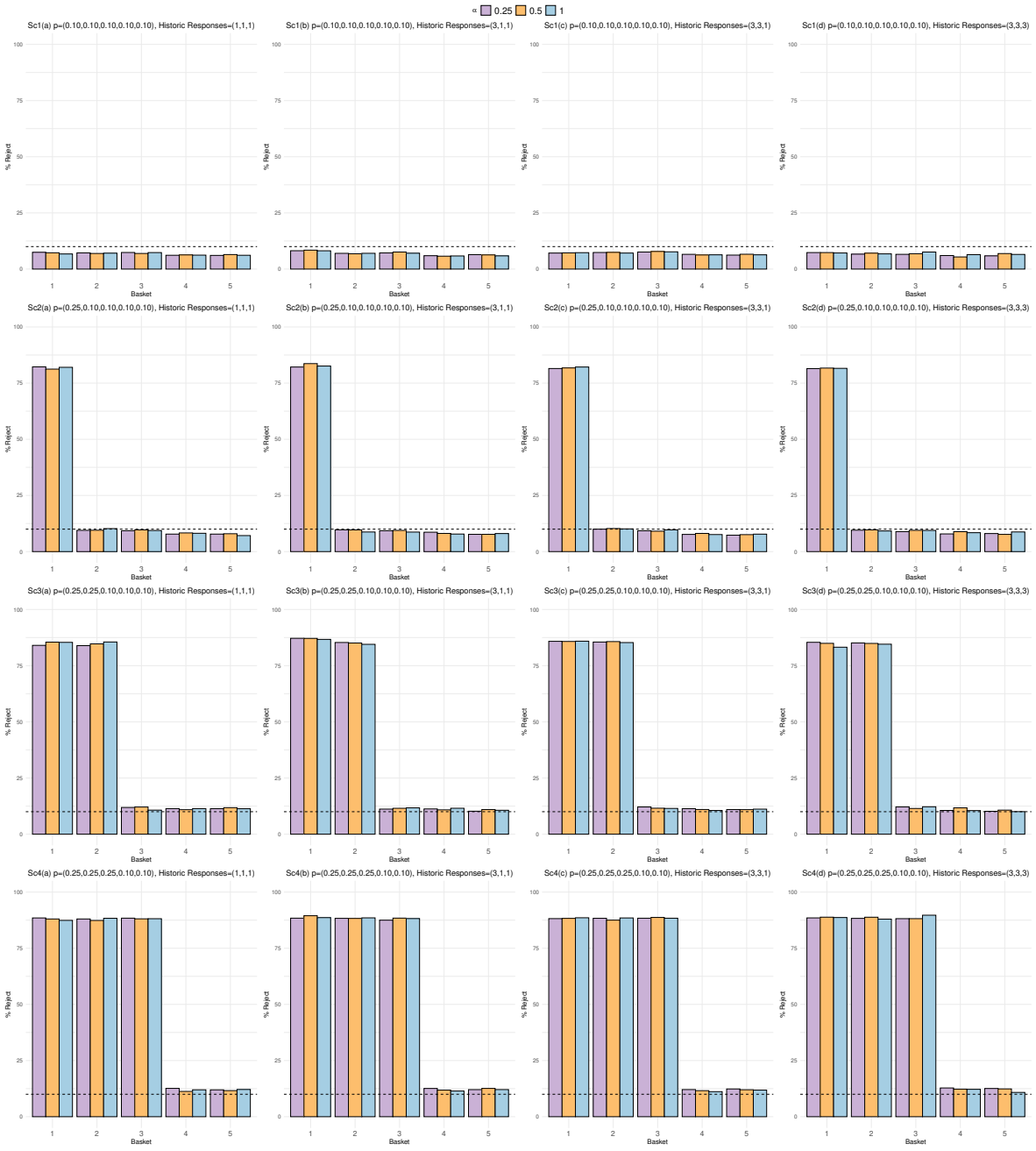


Figure C.5.2: The percentage of data sets where the null hypothesis were rejected per baskets under the EXppNEX model for scenarios 5-8 and 4 historic sub-cases. This is provided for three choices of α: 0.25, 0.5 and 1.

C.6 Exploring the Choice of Weights in the MLMixture Model

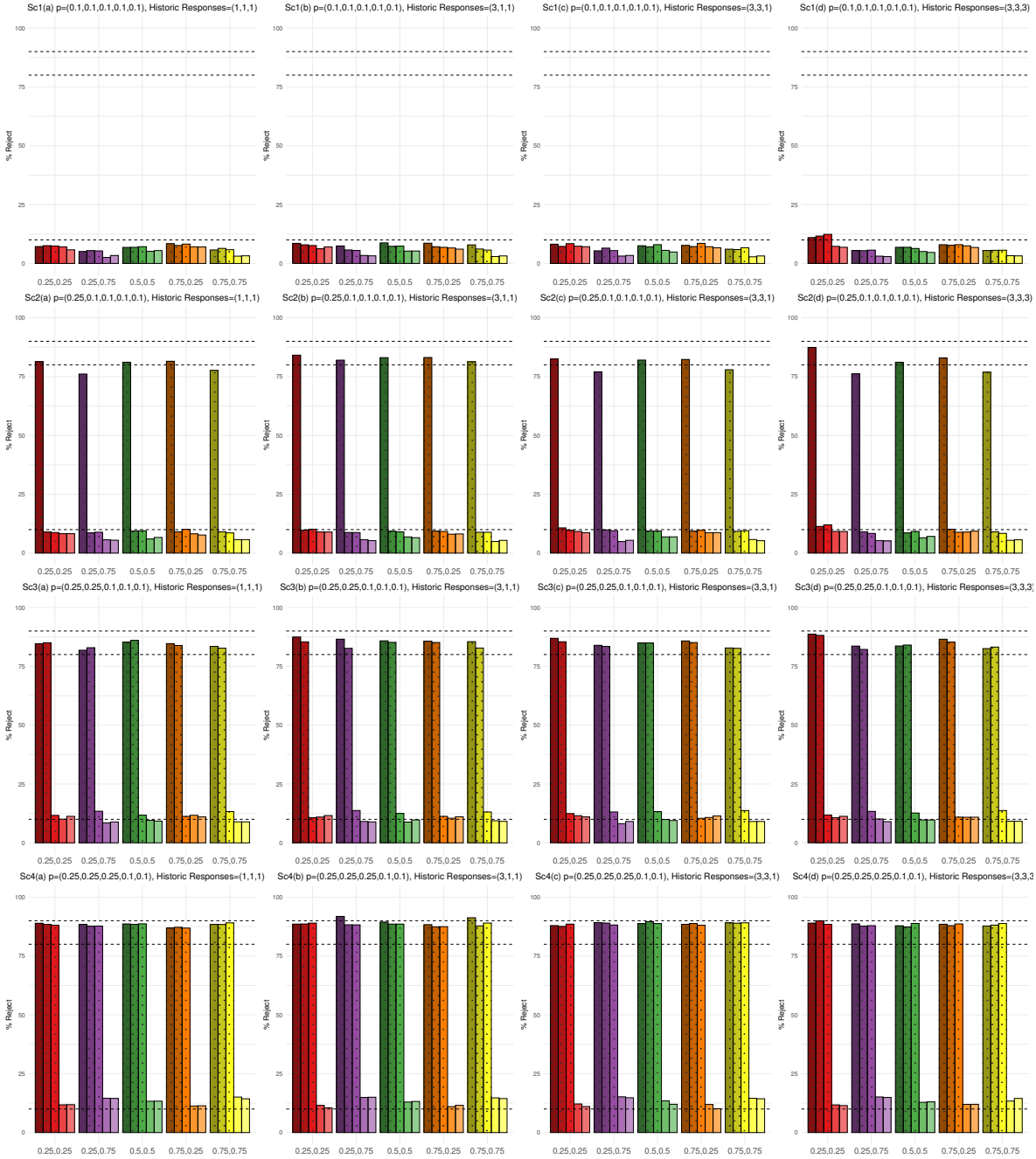


Figure C.6.1: The percentage of data sets where the null hypothesis were rejected per basket under the MLMixture model for scenarios 1-4 and 4 historic sub-cases. This is provided for several choices of $\pi_{\lambda,k}$ and $\pi_{\text{curr},i} = \pi_{\text{all},i}$. Each set of bars labelled x, y correspond to a setting of MLMixture weights where x is the value of $\pi_{\lambda,k}$ (set at 0.25, 0.5 or 0.75) and y are the values of $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$ which are set as equal and to either 0.25, 0.5 or 0.75.

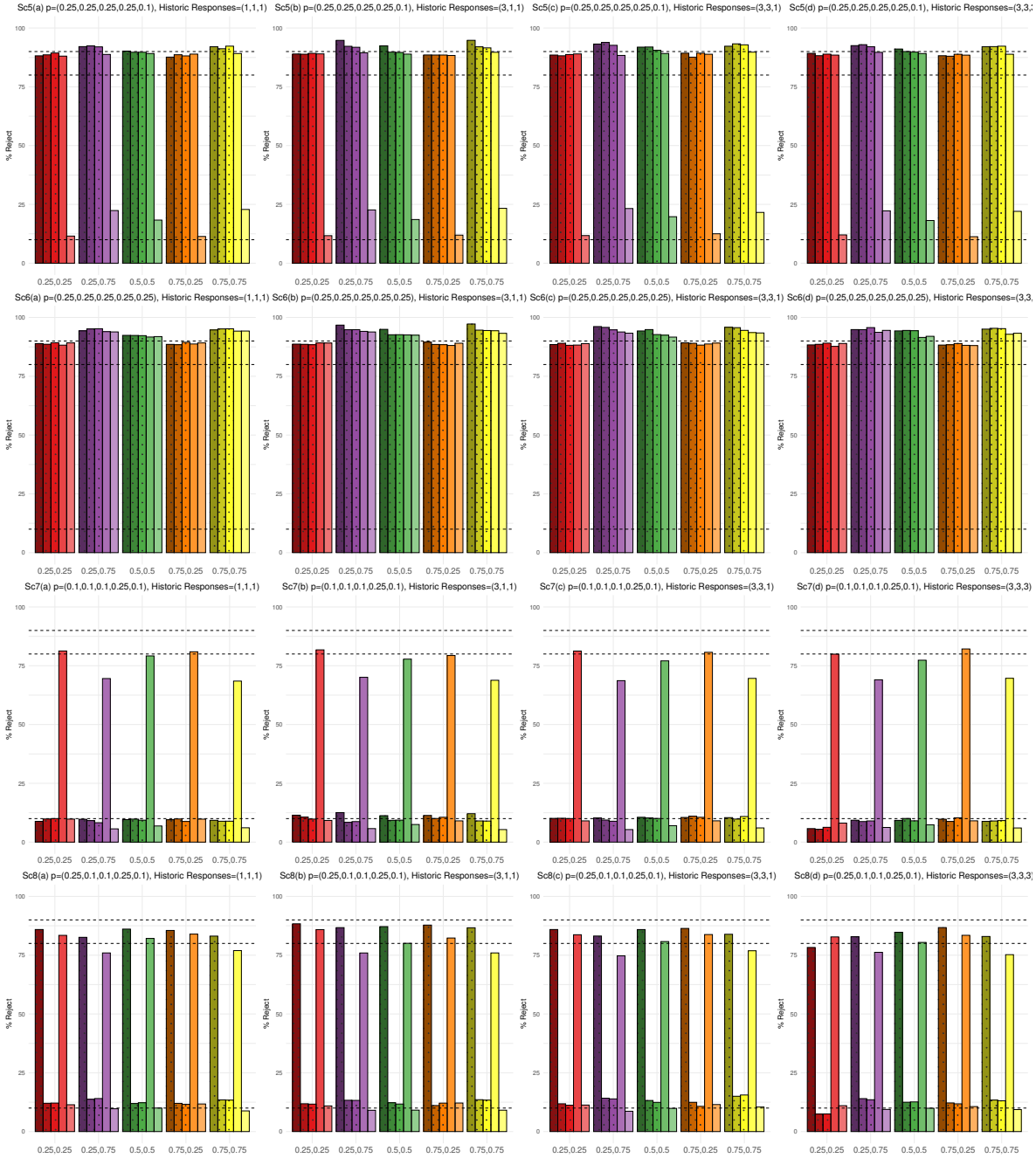


Figure C.6.2: The percentage of data sets where the null hypothesis were rejected per basket under the MLMixture model for scenarios 5-8 and 4 historic sub-cases. This is provided for several choices of $\pi_{\lambda,k}$ and $\pi_{\text{curr},i} = \pi_{\text{all},i}$. Each set of bars labelled x, y correspond to a setting of MLMixture weights where x is the value of $\pi_{\lambda,k}$ (set at 0.25, 0.5 or 0.75) and y are the values of $\pi_{\text{curr},i}$ and $\pi_{\text{all},i}$ which are set as equal and to either 0.25, 0.5 or 0.75.

C.7 Simulation Study with $n_k = 20$ for All Current Baskets, k

The simulation study in Chapter 4 set the sample size of current baskets to be $n_k = 34$ for baskets $k = 1, 2, 3, 4, 5$. This is a particularly large sample size if you compare to the motivating VE-BASKET and MYPathway trials. The large sample size could potentially down-play the benefits of borrowing from the historic information, as baskets with a smaller sample size will benefit more greatly from this additional source of information. To address this, the same simulation study as in Chapter 4 is conducted but the sample size of current baskets reduced to $n_k = 20$ for baskets $k = 1, 2, 3, 4, 5$ (sample size of historic data is still 13 in each). Results are presented in Figures C.7.1-C.7.4.

The comparison between methods holds the same in this study as in the study presented in Chapter 4 with a larger sample size, with performances comparable. Due to the small sample size, the nominal power of 80% is rarely achieved, and in fact, is never achieved using the standard EXNEX model which doesn't consider historical data. The EXppNEX and EXsamNEX also fail to reach this nominal level, however, do get closer. Both mEXNEX$_{\text{hist}}$ and histFujikawa achieve power above 80% under scenarios 5 and 6.

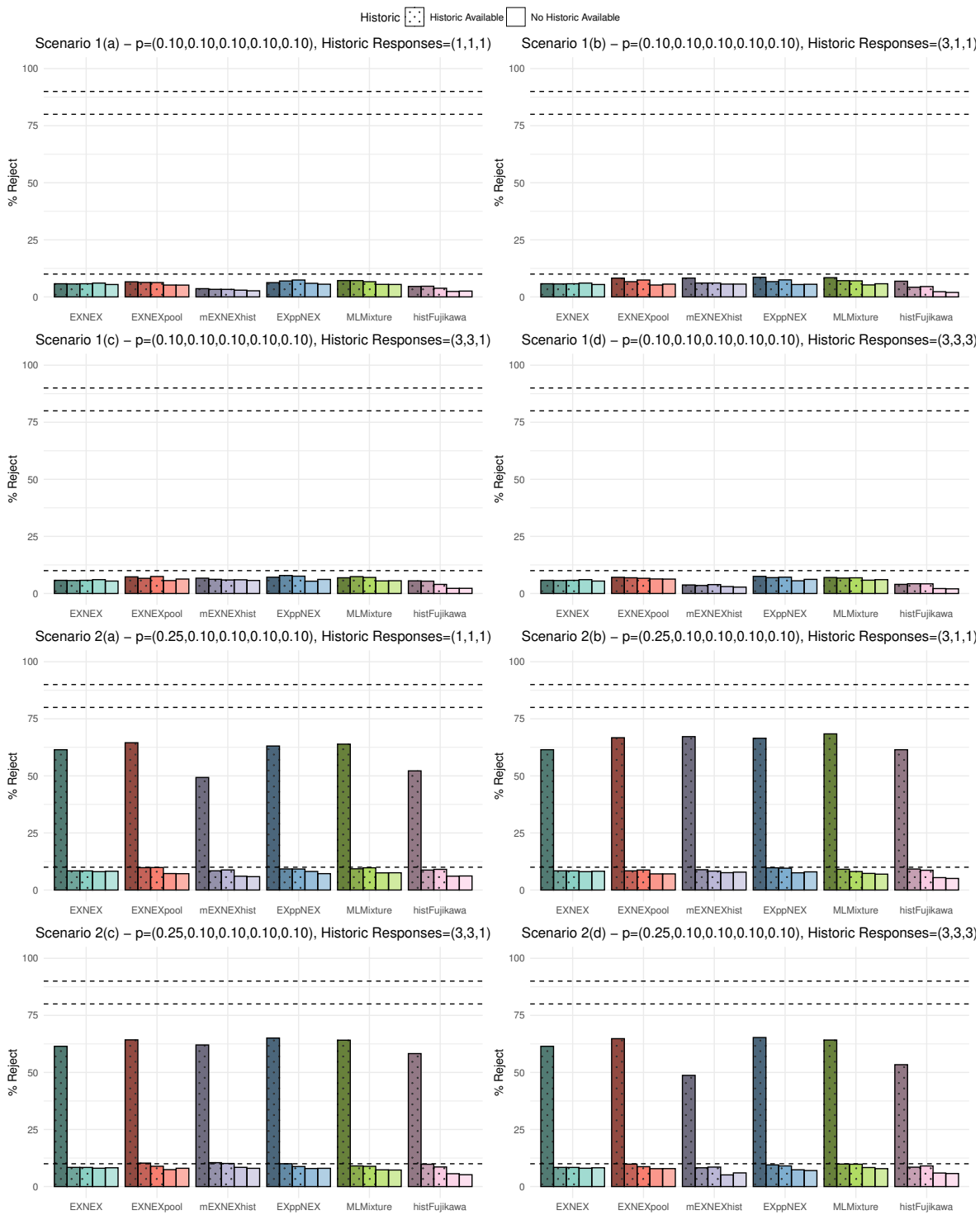


Figure C.7.1: Simulation results for the $n_k = 20$ study: type I error rate and power under each of the 8 approaches for historic information borrowing for scenarios 1 and 2 cases (a)-(d).

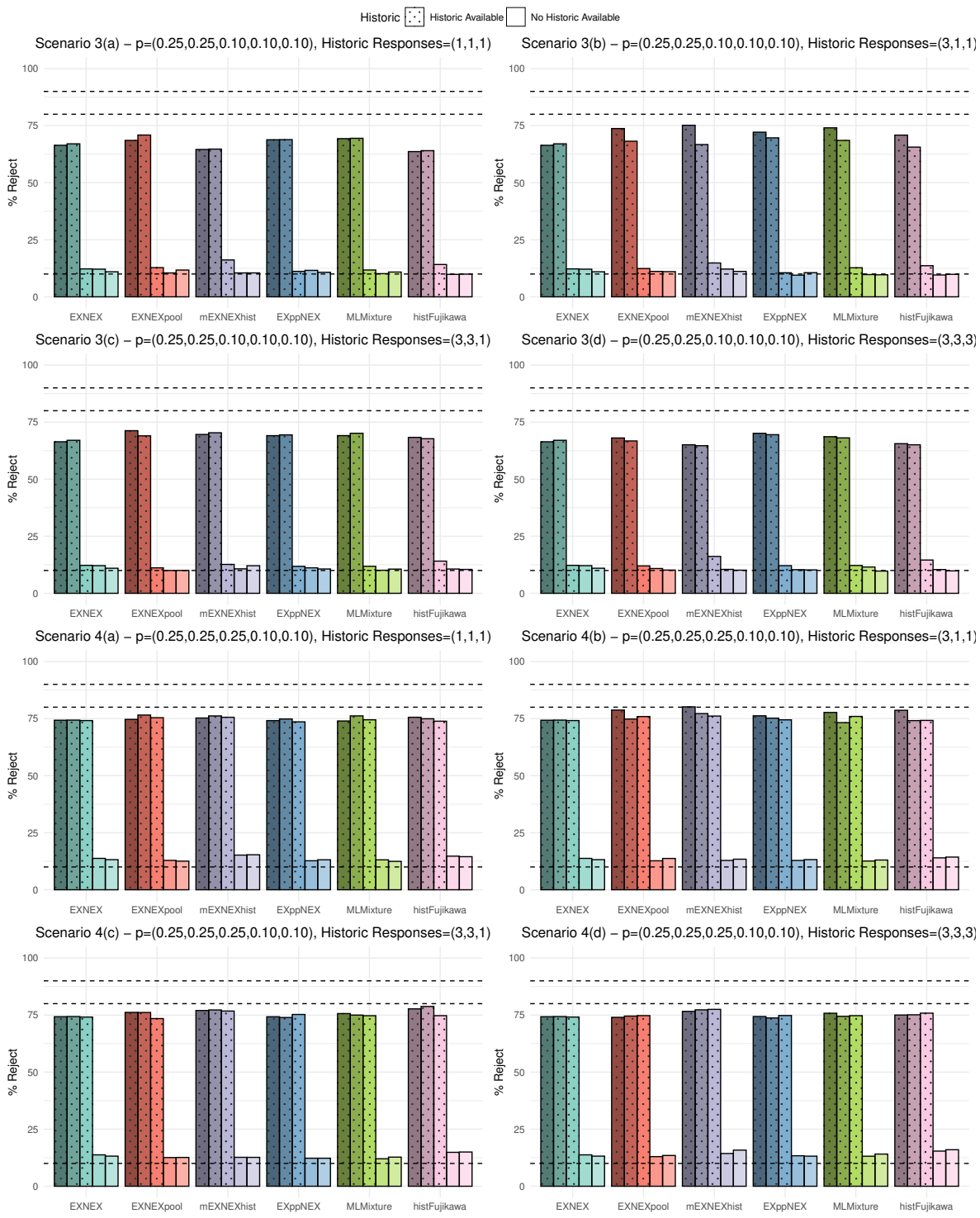


Figure C.7.2: Simulation results for the $n_k = 20$ study: type I error rate and power under each of the 8 approaches for historic information borrowing for scenarios 3 and 4 cases (a)-(d).

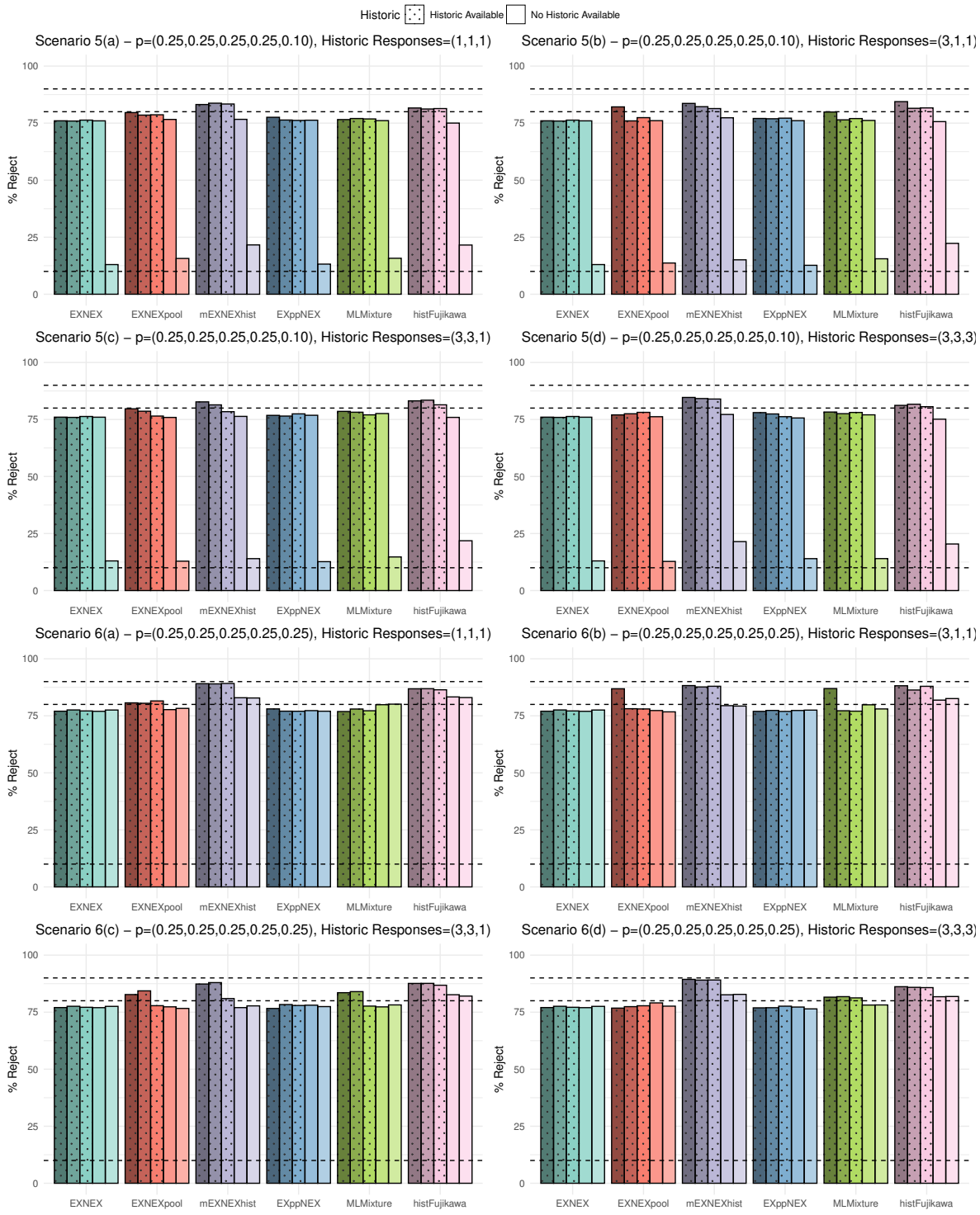


Figure C.7.3: Simulation results for the $n_k = 20$ study: type I error rate and power under each of the 8 approaches for historic information borrowing for scenarios 5 and 6 cases (a)-(d).

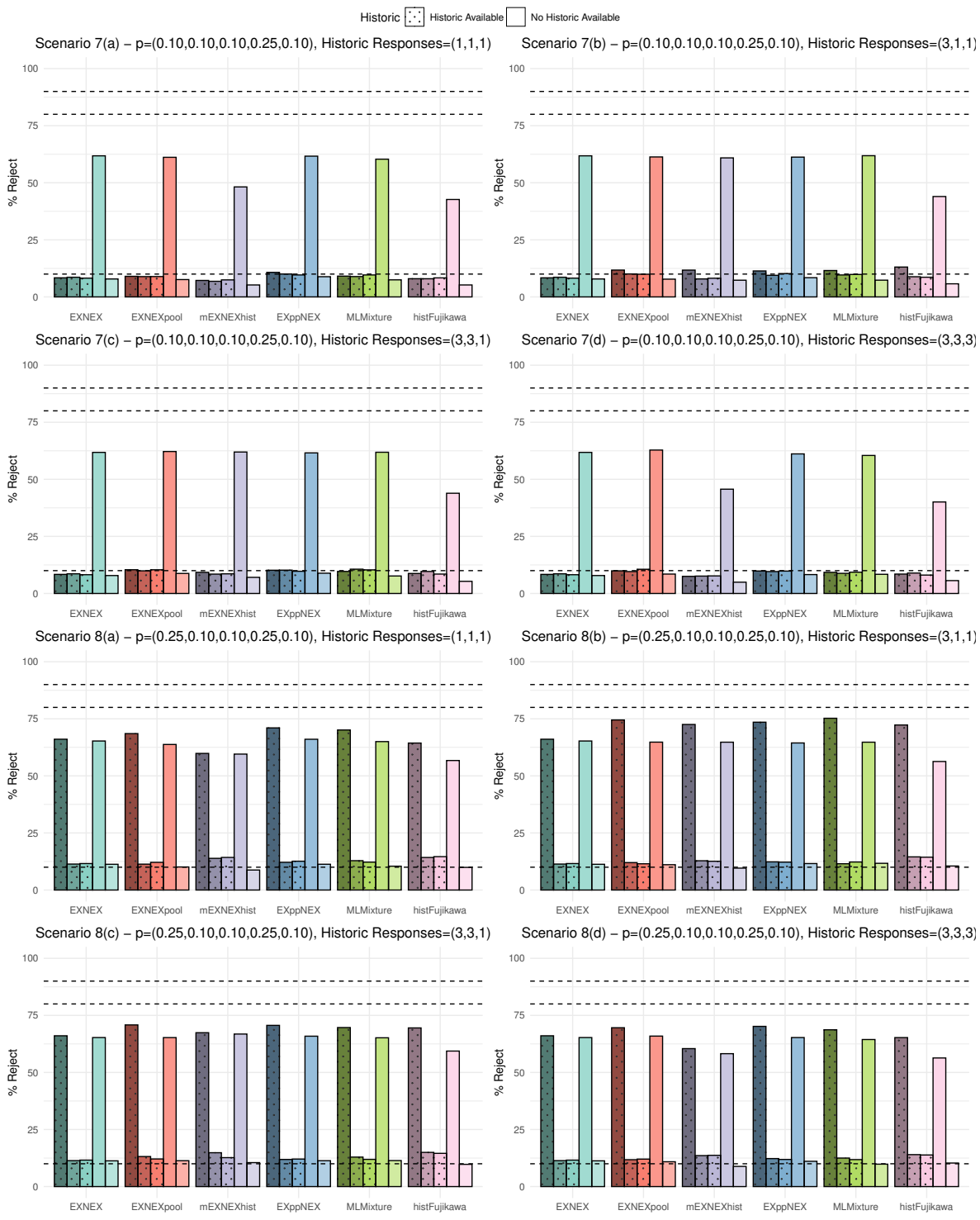


Figure C.7.4: Simulation results for the $n_k = 20$ study: type I error rate and power under each of the 8 approaches for historic information borrowing for scenarios 7 and 8 cases (a)-(d).

Bibliography

Atkinson, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., and Zeger, S. L. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69(3):89–95.

Banbeta, A., van Rosmalen, J., Dejardin, D., and Lesaffre, E. (2019). Modified power prior with multiple historical trials for binary endpoints. *Statistics in Medicine*, 38(7):1147–1169.

Baracaldo-Santamaría, D., Feliciano-Alfonso, J. E., Ramirez-Grueso, R., Rojas-Rodríguez, L. C., Dominguez-Dominguez, C. A., and Calderon-Ospina, C. A. (2023). Making sense of composite endpoints in clinical research. *Journal of Clinical Medicine*, 12(13):4371.

Baumann, L., Sauer, L., and Kieser, M. (2023). Basket trial designs based on power priors. *arXiv preprint arXiv:2309.06988*.

Bennett, M., White, S., Best, N., and Mander, A. (2021). A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: A comparison to standard methods. *Pharmaceutical statistics*, 20(3):462–484.

Bennett, M. S. (2018). *Improving the efficiency of clinical trial designs by using historical control data or adding a treatment arm to an ongoing trial*. PhD thesis, Apollo - University of Cambridge Repository.

Bernardo, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley and Sons.

Berry, S., Broglio, K., Groshen, S., and Berry, D. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials (London, England)*, 10(5):720–734.

Best, N., Ajimi, M., Neuenschwander, B., Saint-Hilary, G., and Wandel, S. (2024). Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors. *Statistics in Biopharmaceutical Research*, 0(0):1–14.

Bhatt, A. (2010). Evolution of clinical research: a history before and beyond James Lind. *Perspectives in clinical research*, 1(1):6–10.

Bogin, V. (2020). Master protocols: new directions in drug discovery. *Contemporary Clinical Trials Communications*, 18:100568.

Bravo, F., Corcoran, T. C., and Long, E. F. (2022). Flexible drug approval policies. *Manufacturing and Service Operations Management*, 24(1).

Chen, C. and Hsiao, C. (2023). Bayesian hierarchical models for adaptive basket trial designs. *Pharmaceutical Statistics*, 22(3):531–546.

Chen, N. and Lee, J. J. (2019). Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes. *Biometrical Journal*, 61(5):1219–1231.

Chen, N. and Lee, J. J. (2020). Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints. *Statistical Methods in Medical Research*, 29(9):2717–2732.

Christyani, G., Carswell, M., Qin, S., and Kim, W. (2024). An overview of advances in rare cancer diagnosis and treatment. *International Journal of Molecular Sciences*, 25(2):1201.

Chu, Y. and Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials (London, England)*, 15(2):149–158.

Cohen, D. R., Todd, S., Gregory, W. M., and Brown, J. M. (2015). Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials*, 16:1–9.

Crofton, J. (2006). The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line. *Journal of the Royal Society of Medicine*, 99(10):531–534.

Cunanan, K., Iasonos, A., Shen, R., and Gönen, M. (2018). Variance prior specification for a basket trial design using Bayesian hierarchical modeling. *Clinical Trials (London, England)*, 16(2):142–153.

Cunanan, K. M., Iasonos, A., Shen, R., Hyman, D. M., Riely, G. J., Gönen, M., and Begg, C. B. (2017). Specifying the true-and false-positive rates in basket trials. *JCO Precision Oncology*, 1.

Daniells, L., Mozgunov, P., Jaki, T., and Bedding, A. (2023). A comparison of Bayesian information borrowing methods in basket trials and a novel proposal of modified exchangeability-nonexchangeability method. *Statistics in Medicine*, 42(24):4392–4417.

Di Liello, R., Piccirillo, M. C., Arenare, L., Gargiulo, P., Schettino, C., Gravina, A., and Perrone, F. (2021). Master protocols for precision medicine in oncology: overcoming methodology of randomized clinical trials. *Life*, 11(11):1253.

Drilon, A., Laetsch, T. W., Kummar, S., DuBois, S. G., Lassen, U. N., Demetri, G. D., Nathenson, M., Doebele, R. C., Farago, A. F., Pappo, A. S., et al. (2018). Efficacy

of larotrectinib in TRK fusion–positive cancers in adults and children. *New England Journal of Medicine*, 378(8):731–739.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics: The Official Journal of the International Environmetrics Society*, 17(1):95–106.

Feinstein, A. R. and Horwitz, R. I. (1982). Double standards, scientific methods, and epidemiologic research. *New england journal of medicine*, 307(26):1611–1617.

Fisher, R. A. (1926). Introduction to the arrangement of field experiments. *J. Minist. Agric. GB*, 33:503–513.

Fleming, T. R., Demets, D. L., and McShane, L. M. (2017). Discussion: The role, position, and function of the fda—the past, present, and future. *Biostatistics*, 18(3):417–421.

Fuglede, B. and Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Fujikawa, K., Teramukai, S., Yokota, I., and Daimon, T. (2020). A Bayesian basket trial design that borrows information across strata based on the similarity between the posterior distributions of the response probability. *Biometrical Journal*, 62(2):330–338.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3).

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., Mei, C., Zhang, L., Deng, C., and Beckman, R. A. (2020). A roadmap to using historical controls in clinical trials–by drug information association adaptive design scientific working group (DIA-ADSWG). *Orphanet journal of rare diseases*, 15:1–19.

Ginsburg, G. S. and Phillips, K. A. (2018). Precision medicine: from science to value. *Health affairs*, 37(5):694–701.

Goetz, L. H. and Schork, N. J. (2018). Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963.

Hainsworth, J. D., Meric-Bernstam, F., Swanton, C., Hurwitz, H., Spigel, D. R., Sweeney, C., Burris, H. A., Bose, R., Yoo, B., Stein, A., Beattie, M., and Kurzrock, R. (2018). Targeted therapy for advanced solid tumors on the basis of molecular profiles: Results from MyPathway, an open-label, phase iia multiple basket study. *Journal of Clinical Oncology*, 36(6):536–542.

Hall, K. T., Vase, L., Tobias, D. K., Dashti, H. T., Vollert, J., Kaptchuk, T. J., and Cook, N. R. (2021). Historical controls in randomized clinical trials: opportunities and challenges. *Clinical Pharmacology & Therapeutics*, 109(2):343–351.

Hariton, E. and Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716.

Hirakawa, A., Asano, J., Sato, H., and Teramukai, S. (2018). Master protocol trials in oncology: review and new trial designs. *Contemporary clinical trials communications*, 12:1–8.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056.

Hobbs, B. P. and Landin, R. (2018). Bayesian basket trial design with exchangeability monitoring. *Statistics in medicine*, 37(25):3557–3572.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: a tutorial (with comments by M. Clyde, D. Draper and E. I. George, and a rejoinder by the authors). *Statistical Science*, 14(4):382–417.

Howard, D. R., Brown, J. M., Todd, S., and Gregory, W. M. (2018). Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Statistical methods in medical research*, 27(5):1513–1530.

Hyman, D., Puzanov, I., Subbiah, V., Faris, J., Chau, I., Blay, J., Wolf, J., Rajae, N., Diamond, E., and Hollebecque, A. (2015). Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine*, 373(8):726–736.

Hyman, D. M., Piha-Paul, S. A., Won, H., Rodon, J., Saura, C., Shapiro, G. I., Juric, D., Quinn, D. I., Moreno, V., Doger, B., et al. (2018). HER kinase inhibition in patients with HER2-and HER3-mutant cancers. *Nature*, 554(7691):189–194.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Jin, J., Riviere, M., Luo, X., and Dong, Y. (2020). Bayesian methods for the analysis of early-phase oncology basket trials with information borrowing across cancer types. *Statistics in Medicine*, 39(25):3459–3475.

Jones, D. S. and Podolsky, S. H. (2015). The history and fate of the gold standard. *The Lancet*, 385(9977):1502–1503.

Kaizer, A., Zabor, E., Nie, L., and Hobbs, B. (2022). Bayesian and frequentist approaches to sequential monitoring for futility in oncology basket trials: A comparison of simon's two-stage design and Bayesian predictive probability monitoring with information sharing across baskets. *PloS one*, 17(8):e0272367–e0272367.

Kaizer, A. M., Belli, H. M., Ma, Z., Nicklawsky, A. G., Roberts, S. C., Wild, J., Wogu, A. F., Xiao, M., and Sabo, R. T. (2023a). Recent innovations in adaptive trial designs: a review of design opportunities in translational research. *Journal of Clinical and Translational Science*, pages 1–35.

Kaizer, A. M., Belli, H. M., Ma, Z., Nicklawsky, A. G., Roberts, S. C., Wild, J., Wogu, A. F., Xiao, M., and Sabo, R. T. (2023b). Recent innovations in adaptive trial designs: a review of design opportunities in translational research. *Journal of Clinical and Translational Science*, 7(1):e125.

Kemp, R. and Prasad, V. (2017). Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC medicine*, 15:1–7.

Kopp-Schneider, A., Calderazzo, S., and Wiesenfarth, M. (2020). Power gains by using external information in clinical trials are typically not possible when requiring strict type i error control. *Biometrical Journal*, 62(2):361–374.

Korn, E. L. and Freidlin, B. (2017). Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. *JNCI: Journal of the National Cancer Institute*, 109(6):djx013.

Lang, T. (2011). Adaptive trial design: could we use this approach to improve clinical trials in the field of global health? *The American journal of tropical medicine and hygiene*, 85(6):967.

Lee, S. M. and Cheung, Y. K. (2009). Model calibration in the continual reassessment method. *Clinical Trials*, 6(3):227–238.

Lee, S. M. and Cheung, Y. K. (2011). Calibration of prior variance in the Bayesian continual reassessment method. *Statistics in Medicine*, 30(17):2081–2089.

Liu, R., Liu, Z., Ghadessi, M., and Vonk, R. (2017). Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemporary clinical trials*, 63:67–72.

Lu, C., Li, X., Broglio, K., Bycott, P., Jiang, Q., Li, X., McGlothlin, A., Tian, H., and Ye, J. (2021). Practical considerations and recommendations for master protocol framework: basket, umbrella and platform trials. *Therapeutic Innovation & Regulatory Science*, 55(6):1145–1154.

Lyu, J., Zhou, T., Yuan, S., Guo, W., and Ji, Y. (2023). MUCE: Bayesian hierarchical modelling for the design and analysis of phase 1b multiple expansion cohort trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(3):649–669.

Mano, H., Tanaka, Y., Orihara, S., and Moriya, J. (2023). Application of sample size re-estimation in clinical trials: A systematic review. *Contemporary Clinical Trials Communications*, page 101210.

Marion, J. D. and Althouse, A. D. (2023). The use of historical controls in clinical trials. *JAMA*, 330(15):1484–1485.

Meadows, M. (2006). Promoting safe and effective drugs for 100 years. *FDA Consumer magazine*, 40(1).

Muehlemann, N., Zhou, T., Mukherjee, R., Hossain, M. I., Roychoudhury, S., and Russek-Cohen, E. (2023). A tutorial on modern Bayesian methods in clinical trials. *Therapeutic Innovation & Regulatory Science*, 57(3):402–416.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics : the Journal of the Pharmaceutical Industry*, 15(2):123–134.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Schmidli, H. (2023). On fixed and uncertain mixture prior weights. *arXiv preprint arXiv:2306.15197*.

Oakes, J. (2013). Effect identification in comparative effectiveness research. *EGEMS (Washington DC)*, 1(1):1004–1004.

Okeke, F., Nriagu, V. C., Nwaneki, C. M., Magacha, H. M., Omenuko, N. J., and Anazor, S. (2023). Factors that determine multiple primary cancers in the adult population in the united states. *Cureus*, 15(9).

Ouma, L., Grayling, M., Wason, J., and Zheng, H. (2022a). Bayesian modelling strategies for borrowing of information in randomised basket trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):2014–2037.

Ouma, L. O., Wason, J. M., Zheng, H., Wilson, N., and Grayling, M. (2022b). Design and analysis of umbrella trials: Where do we stand? *Frontiers in Medicine*, 9:1037439.

Pan, H., Yuan, Y., and Xia, J. (2017). A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(5):979–996.

Park, J. J., Hsu, G., Siden, E. G., Thorlund, K., and Mills, E. J. (2020). An overview of precision oncology basket and umbrella trials for clinicians. *CA: a cancer journal for clinicians*, 70(2):125–137.

Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K., and Mills, E. J. (2019). Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*, 20:1–10.

Pessoa-Amorim, G., Campbell, M., Fletcher, L., Horby, P., Landray, M., Mafham, M., and Haynes, R. (2021). Making trials part of good clinical care: lessons from the recovery trial. *Future Healthcare Journal*, 8(2):e243–e250.

Plummer, M. (2023). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-15.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188.

Psioda, M., Xu, J., Jinag, Q., Ke, C., Yang, Z., and Ibrahim, J. (2021). Bayesian adaptive basket trial design using model averaging. *Biostatistics (Oxford, England)*, 22(1):19–34.

Psioda, M. A. and Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20(3):400–415.

Qin, B.-D., Jiao, X.-D., Liu, K., Wu, Y., He, X., Liu, J., Qin, W.-X., Wang, Z., and Zang, Y.-S. (2019). Basket trials for intractable cancer. *Frontiers in Oncology*, 9:229.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reck, M., Okines, A., Pohlmann, P., Yu, E., Bekaii-Saab, T., Nakamura, Y., Monk, B., O'Malley, D., Kang, V., Walker, L., et al. (2021). 557TiP SGNTUC-019: phase II basket study of tucatinib and trastuzumab in previously treated solid tumors with HER2 alterations. *Annals of Oncology*, 32:S614–S615.

Rosier, J. A., Martens, M. A., and Thomas, J. R. (2014). *Drug Life Cycle*, chapter 1, pages 1–11. John Wiley and Sons, Ltd.

Sargent, D. and Renfro, L. (2017). Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of Oncology*, 28(1):34–43.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

Sedgwick, P. (2011). Phases of clinical trials. *Bmj*, 343.

Shahapur, P. R., Vadakedath, S., Bharadwaj, V. G., Kumar, P., Pinnelli, V. B., Godishala, V., Kandi, V., et al. (2022). Research question, objectives, and endpoints in clinical and oncological research: a comprehensive review. *Cureus*, 14(9).

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10.

Simon, R. and Roychowdhury, S. (2013). Implementing personalized cancer genomics in clinical trials. *Nature Publishing Group*, 12(5):358–369.

Steinmetz, K. L. and Spack, E. G. (2009). The basics of preclinical drug development for neurodegenerative disease indications. *BMC neurology*, 9(Suppl 1):S2.

Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.

Streiner, D. L. (2015). Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. *The American journal of clinical nutrition*, 102(4):721–728.

Strzebonska, K. and Waligora, M. (2019). Umbrella and basket trials in oncology: ethical challenges. *BMC medical ethics*, 20:1–10.

Su, L., Chen, X., Zhang, J., and Yan, F. (2022). Comparative study of Bayesian information borrowing methods in oncology clinical trials. *JCO Precision Oncology*, 6:e2100394.

Tao, J., Schram, A., and Hyman, D. (2018). Basket studies: Redefining clinical trials in the era of genome-driven oncology. *Annual Review of Medicine*, 69(1):319–331.

Turner, J. R. (2010). *New drug development: An introduction to clinical trials: Second edition.* Springer.

U.S Food and Drug Administration (2023a). Considerations for the design and conduct of externally controlled trials for drug and biological products.

U.S Food and Drug Administration (2023b). Master protocols for drug and biological product development.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., and Lesaffre, E. (2018). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical methods in medical research*, 27(10):3167–3182.

Vogt, A., Schmid, S., Heinimann, K., Frick, H., Herrmann, C., Cerny, T., and Omlin, A. (2017). Multiple primary tumours: challenges and approaches, a review. *ESMO open*, 2(2):e000172.

Wang, H. and Yee, D. (2019). I-SPY 2: a neoadjuvant adaptive clinical trial designed to improve outcomes in high-risk breast cancer. *Current breast cancer reports*, 11:303–310.

Weber, S., Li, Y., Seaman, J. W., Kakizume, T., and Schmidli, H. (2021). Applying meta-analytic-predictive priors with the R bayesian evidence synthesis tools. *Journal of Statistical Software*, 100.

Wojciekowski, S. (2022). *bhmbasket: Bayesian Hierarchical Models for Basket Trials*. R package version 0.9.5.

Woodcock, J. and LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1):62–70.

Yamaguchi, S., Kaneko, M., and Narukawa, M. (2021). Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clinical and Translational Science*, 14(3):1113–1122.

Yang, P., Zhao, Y., Nie, L., Vallejo, J., and Yuan, Y. (2023). SAM: Self-adapting mixture prior to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79(4):2857–2868.

Zhang, H., Shen, Y., Chiang, A. Y., and Li, J. (2021). An empirical Bayes robust meta-analytical-predictive prior to adaptively leverage external data. *arXiv preprint arXiv:2109.10237*.

Zheng, H., Grayling, M. J., Mozgunov, P., Jaki, T., and Wason, J. M. (2023). Bayesian sample size determination in basket trials borrowing information between subsets. *Biostatistics*, 24(4):1000–1016.

Zheng, H. and Hampson, L. V. (2020). A Bayesian decision-theoretic approach to incorporate preclinical information into phase I oncology trials. *Biometrical Journal*, 62(6):1408–1427.

Zheng, H. and Wason, J. M. (2022). Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics*, 23(1):120–135.

Zhou, T. and Ji, Y. (2020). RoBoT: a robust Bayesian hypothesis testing method for basket trials. *Biostatistics*, 22(4):897–912.