Aus dem Institut für Medizinische Biometrie
Universitätsklinikum Heidelberg
(Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser)

A Computationally Efficient Basket Trial Design Based on Power Priors

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Lukas Baumann
aus
Wien, Österreich

2024

Dekan:	Herr Prof. Dr. Michael Boutros
Doktorvater:	Herr Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations

BHM	Bayesian hierarchical model
BMA	Bayesian model averaging
CPP	calibrated power prior
CRAN	comprehensive R archive network
ECD	expected number of correct decisions
EXNEX	exchangeability-nonexchangeability
FWER	family-wise type 1 error rate
KLD	Kullback-Leibler divergence
KS	Kolmogorov-Smirnov
JSD	Jensen-Shannon divergence
MCMC	Markov Chain Monte Carlo
MEM	multisource exchangeability model
MML	maximum marginal likelihood
TOER	type 1 error rate

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Historically, the tissue of origin of a tumour was the decisive factor for how cancer was treated and for how cancer drugs were developed (Redig and Jänne, 2015; Jørgensen, 2019). In recent years, however, the availability of new genomic tools and a better understanding of cancer pathogenesis have led to new cancer therapies that do not focus on tumour histology or location but target tumours with a certain genetic alteration (Simon, 2018; Fridlyand et al., 2013). As a consequence of this paradigm shift in cancer drug development, new clinical trial designs were developed, as trials that only include patients with a certain tumour histology or location may become infeasible when the targeted genetic alteration is rare. So-called basket trials address this issue by including different patient subgroups which are called baskets. In oncology, basket trials usually include patients with different primary tumour types, which form the baskets, that all share a common feature, such as a genetic mutation (Hirakawa et al., 2018). They are typically single-arm phase II trials with tumour response as the binary endpoint. Many of these trials analyse each basket individually or pool the data of all baskets for the analysis (Hobbs et al., 2022). However, a separate analysis of each basket has no statistical benefits compared to separate studies for each subgroup, hence also leads to low power when subgroups are small. A pooled analysis may increase the power, but is only a sensible strategy when efficacy is similar in all patient subgroups. However, this is not

always the case. For example, a basket trial investigating vemurafenib showed response rates of more than 40% in some subgroups, but response rates of less than 15% in other subgroups of patients with a BRAF V600 mutation (Hyman et al., 2015). Hence, the challenge in analysing basket trials is to use techniques that lead to higher power than a separate analysis in each subgroup but combine the available data in a more nuanced way than by a simple pooled analysis, to reliably identify the subgroups in which the treatment is effective.

1.2 Previous Work

A wide range of different designs for the analysis of basket trials using both frequentist and Bayesian methodology have been suggested in recent years (Pohl et al., 2021). In the frequentist design proposed by Cunanan et al. (2017), for example, a test for heterogeneity is performed to decide whether the baskets are pooled or analysed individually. Similarly, in the pruning-and-pooling design, baskets that reach a certain efficacy threshold are pooled for the final analysis, whereas inactive baskets are excluded from the final analysis or already stopped in an interim analysis (Chen et al., 2016, 2021; Zhou et al., 2019). Another frequentist design was suggested by Krajewska and Rauch (2021), where k-means clustering is used to identify clusters of baskets that respond similarly to the treatment. Baskets within a cluster are then pooled for the analysis.

Bayesian methods are, however, more common as they have the advantage that sharing information between subgroups is possible in more sophisticated ways than by simply pooling them and is a feature of many Bayesian models. The amount of data that is shared depends on the observed data and pre-specified prior or other tuning parameters. Many designs utilise Bayesian hierarchical modelling that assumes a common distribution of the (transformed) response probabilities of all baskets. In the easiest case, the logit-transformed response probabilities are assumed to arise from a single normal distribution (Thall et al., 2003; Berry et al., 2013). Neuenschwander et al. (2016) extended this idea by fitting both a Bayesian hierarchical model to all baskets and an individual model separately for each basket. These two models are then combined with a pre-specified weight. Another Bayesian tool that is used for analysing basket trials is Bayesian model averaging. In the design of Psioda et al. (2021), all possible cluster configurations of baskets are constructed, ranging from single-basket clusters

to one cluster that contains all baskets. Within the clusters all data are pooled and analysed using a beta-binomial model. The different cluster configurations are then weighted by their posterior probabilities.

Bayesian hierarchical modelling and Bayesian model averaging are computationally relatively intense. The posterior distributions for response probabilities in the designs based on Bayesian hierarchical modelling are not available in closed form. Therefore, Markov Chain Monte Carlo (MCMC) sampling is necessary to obtain posterior probabilities (Thall et al., 2003). In the Bayesian model averaging design, the posterior distributions can be derived analytically, but as the number of baskets increases, the number of clusters and therefore the computational burden increase exponentially.

Another design that uses Bayesian tools was proposed by Fujikawa et al. (2020), which is henceforth called Fujikawa's design. In this design, at first a beta-binomial model is applied to analyse each basket individually. The similarity between individual baskets is then quantified by computing the Jensen-Shannon divergence (JSD) of all pairs of posterior distributions that arise from the beta-binomial models. Weights between 0 and 1 that determine the amount of data that is shared between two baskets are then computed based on the JSD and two tuning parameters. Beta-binomial models are then used again, where the posterior parameters are computed as weighted sums of the posterior parameters of the individual models. Since the weights are solely derived from the available data and no prior distribution is specified for them, Fujikawa's design can be considered an empirical Bayes design. Since the posteriors are beta distributions, the posterior probabilities can be calculated without using MCMC sampling. The weights necessary to compute the posterior parameters can also be computed fast using numerical integration. Thus, Fujikawa's design is computationally much cheaper than other (fully) Bayesian designs. However, there is currently no comprehensive simulation study that compares the performance of Fujikawa's design to other basket trial designs and how the performance of the design is influenced by the choice of tuning parameter values. A further open question is whether Fujikawa's design can be improved by using other measures than the JSD to derive the weights.

1.3 Objective and Structure of the Thesis

The main objective of the thesis is to extend Fujikawa's design and thereby improve its performance in terms of relevant operating characteristics and enable a better fine tuning of the amount of shared information. Specifically, while this was not noted in their manuscript, Fujikawa's design is closely related to the concept of power priors, that was initially proposed to use historical data in a prospective clinical trial (Ibrahim and Chen, 2000). This connection is established in the thesis and different methods from the power prior literature are adapted to basket trials. It is investigated how Fujikawa's design and the newly proposed modifications perform under different scenarios compared to other Bayesian basket trial designs. A further objective is to investigate nonmonotonicity of test decisions in the number of observed responses. Nonmonotonic events are defined as outcomes for which a null hypothesis can be rejected, while another outcome with a higher observed response rate exists that does not lead to a rejected null hypothesis. These events can arise as a consequence of data-dependent information sharing and have not been discussed in the literature in the context of basket trials before.

The structure of the thesis is as follows: In Chapter 2, all methods used in the thesis are described in detail. This chapter starts with the basic setup and notation in Section 2.1 and continues with the description of some essential Bayesian tools that are used in basket trial designs in Section 2.2. The JSD is introduced in Section 2.3. In Section 2.4, Fujikawa's design and other basket trial designs used for comparison are explained. The operating characteristics used to compare the different designs are defined in Section 2.5. The concept of power priors and the specific methods later adapted for basket trials are introduced in Section 2.6. In Chapter 3, the results of the thesis are presented. In Section 3.1, it is demonstrated how power priors can be used to share information in basket trials and how this is connected to Fujikawa's design. The setup and results of a comparison study are shown in Section 3.2. In Section 3.3, nonmonotonicity of test decisions is investigated. The R packages baskexact and basksim in which all newly proposed methods are implemented are presented in Section 3.4 and Section 3.5. The thesis concludes with a discussion in Chapter 4.

Chapter 2

Methods

2.1 Setup and Notation

Throughout the thesis an uncontrolled single-arm basket trial is considered. A treatment is investigated in K disjoint subgroups, which are called baskets. The sample size in each basket is denoted by n_k, with $k \in \{1, ..., K\}$. The observations of a given basket are considered exchangeable. The objective of the trial is to evaluate in which baskets the treatment is effective. The endpoint is binary, the number of responses in basket k, denoted by r_k, are a realisation from a random variable. The vector of realisations is denoted by $\boldsymbol{r} = (r_1, \dots r_K)$. $p_k \in [0, 1]$ denotes the response probability in basket k. Note that throughout the thesis the term "probability" refers to a true and in practice unknown quantity, i.e. the estimand, while the estimates for these values are referred to as "rate".

Although basket trials are often evaluated using Bayesian tools, and all methods considered in this thesis have at least some Bayesian elements, a (frequentist) hypothesis pair is often defined. Since the considered trials are single-arm, the hypotheses include a fixed response probability $p_0 \in (0, 1)$ which is for example derived from the estimated response probability of the standard of care. While the null response probability may in general differ between baskets, throughout the thesis it is assumed that p_0 is equal for all baskets. The treatment under investigation is considered of clinical interest if the response probability is larger than

p_0. Hence, the hypotheses to be tested are:

$$H_{0,k} : p_k \leqslant p_0 \text{ vs. } H_{1,k} : p_k > p_0 \tag{2.1}$$

for $k \in \{1, \ldots, K\}$.

2.2 Basic Bayesian Tools

In this section, some basic Bayesian methods are explained, which are necessary to better understand the basket trial designs introduced in Section 2.4. The Bayesian tools introduced here are applied in a wide range of areas in Bayesian statistics and are not specific to basket trials. Therefore, they are also introduced here in more general terms. This is especially important to emphasise to avoid confusion between the concept of Bayesian hierarchical modelling (see Section 2.2.2) and the BHM basket trial design by Berry et al. (2013) (see Section 2.4.4) as well as the concept of Bayesian model averaging (see Section 2.2.3) and the BMA basket trial design by Psioda et al. (2021) (see Section 2.4.3) as in this thesis these two designs are named after the Bayesian method they utilise. To distinguish between general methods and basket trial designs, the abbreviations BMA and BHM are only used when referring to the respective design.

2.2.1 Beta-Binomial Model

The beta-binomial model (e.g. Gelman et al., 2004, p. 33-34) is used to model the response probability p of a binomially distributed variable with realisation r, that arises from the sum of n exchangeable Bernoulli experiments, i.e. experiments with two possible outcomes: a success or a non-success. The likelihood of p given r is:

$$L(p|r) = \mathbb{P}(r|p) = \binom{n}{r} \cdot p^r \cdot (1-p)^{n-r}. \tag{2.2}$$

Note that $\mathbb{P}(r|p)$ could more formally be written as $\mathbb{P}_p(R = r)$, where R denotes the random variable that r arises from. In the classical statistical literature, parameters of the distribution are usually written as a subscript of $\mathbb{P}$ to emphasise that they are fixed values and not random

variables, but in the Bayesian literature the notation $\mathbb{P}(r|p)$ is more common and therefore also used in the following.

For p, a beta prior distribution with shape parameters $s_1 > 0$ and $s_2 > 0$, denoted by $\pi_0(p) = \text{Beta}(s_1, s_2)$ is specified. The two parameters of the beta distribution correspond to the number of successes r and non-successes $n - r$ as is seen in Equation (2.2). Thus, if for example $s_1 = s_2 = 1$, the beta prior contains as much information as observing one success and one non-success.

The density of a beta distribution is (e.g. Christensen et al., 2011, p. 100):

$$f(p) = \frac{1}{\mathcal{B}(s_1, s_2)} \cdot p^{s_1-1} \cdot (1-p)^{s_2-1},$$

where $\mathcal{B}(\cdot, \cdot)$ is the beta function, which is defined as (e.g. Kruschke, 2015, p. 127):

$$\mathcal{B}(x, y) = \int_0^1 t^{x-1} \cdot (1-t)^{y-1} dt.$$

The posterior distribution of p given r, denoted by $\pi(p|r)$, has a closed-form solution, since the beta prior is conjugate for the binomial likelihood (e.g. Bernardo and Smith, 2000, p. 267):

$$\begin{aligned}
\pi(p|r) &\propto L(p|r) \cdot \pi_0(p) \\
&= \binom{n}{r} \cdot p^r \cdot (1-p)^{n-r} \cdot \frac{1}{\mathcal{B}(s_1, s_2)} \cdot p^{s_1-1} \cdot (1-p)^{s_2-1} \\
&\propto p^r \cdot (1-p)^{n-r} \cdot p^{s_1-1} \cdot (1-p)^{s_2-1} \\
&= p^{s_1+r-1} \cdot (1-p)^{s_2+n-r-1} \\
&\propto \text{Beta}(s_1 + r, s_2 + n - r). \qquad (2.2)
\end{aligned}$$

Hence, the posterior is a beta distribution and the shape parameters are found by adding the number of successes r to the first prior shape parameter s_1 and the number of non-successes $n - r$ to the second prior shape parameter s_2.

The posterior predictive distribution is of interest to compute the probability of future observations. This distribution has a closed-form solution for the beta-binomial model (e.g.

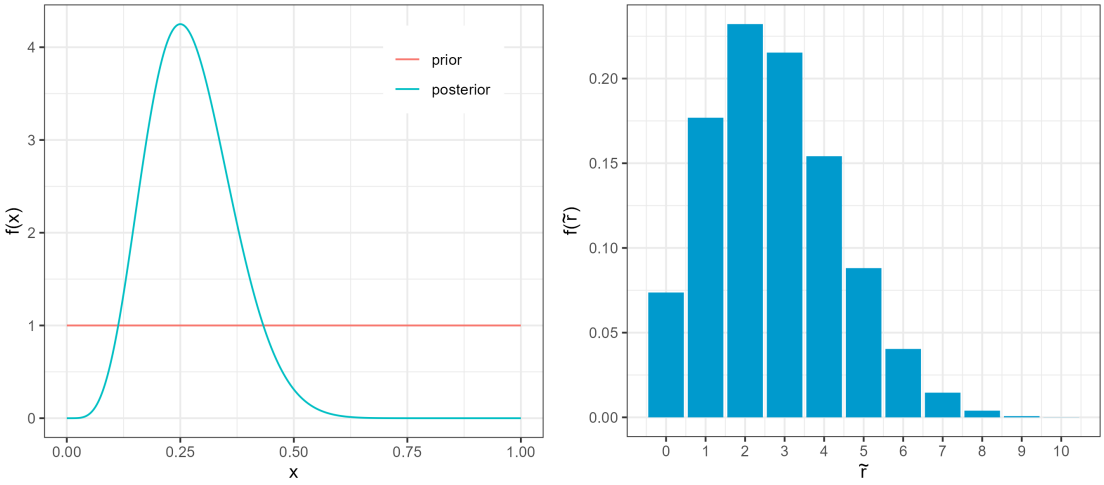

Figure 1: *Prior and posterior (left-hand side) and posterior predictive distribution (right-hand side) of a beta-binomial model. The prior distribution is Beta(1, 1) and the posterior distribution is Beta (6, 16) which results from 5 successes in 20 Bernoulli experiments. The posterior predictive distribution for a future observation is based on $m = 10$ Bernoulli experiments given the observed data and the prior distribution.*

Christensen et al., 2011, p. 25). Let $\tilde{r}$ be a future binomial observation, based on m Bernoulli experiments, which is conditionally independent of r given p. With the posterior predictive distribution one can compute the probability of observing $\tilde{r}$ out of m successes, given the observed data r and the prior distribution:

$$\begin{aligned}
\mathbb{P}(\tilde{r}|r) &= \int_0^1 \mathbb{P}(\tilde{r}|r,p) \cdot \pi(p|r)dp \\
&= \int_0^1 \mathbb{P}(\tilde{r}|p) \cdot \pi(p|r)dp \quad \text{(conditional independence)} \\
&= \int_0^1 \binom{m}{\tilde{r}} \cdot p^{\tilde{r}} \cdot (1-p)^{m-\tilde{r}} \cdot \frac{1}{\mathcal{B}(s_1+r, s_2+n-r)} \cdot p^{s_1+r-1} \cdot (1-p)^{s_2+n-r-1} dp \\
&= \binom{m}{\tilde{r}} \cdot \frac{1}{\mathcal{B}(s_1+r, s_2+n-r)} \cdot \int_0^1 p^{s_1+r+\tilde{r}-1} \cdot (1-p)^{s_2+n-r+m-\tilde{r}-1} dp \\
&= \binom{m}{\tilde{r}} \cdot \frac{\mathcal{B}(s_1+r+\tilde{r}, s_2+n-r+m-\tilde{r})}{\mathcal{B}(s_1+r, s_2+n-r)},
\end{aligned} \tag{2.3}$$

which is the density of a beta-binomial distribution with parameters $s_1 + r$, $s_2 + n - r$ and m (e.g. Bernardo and Smith, 2000, p. 117).

Figure 1 illustrates the prior, posterior and posterior predictive distribution with an example. A Beta$(1, 1)$ prior distribution is used and 20 Bernoulli experiments resulted in 5 successes.

The posterior predictive distribution based on the prior and the observed data is shown for $m = 10$.

2.2.2 Bayesian Hierarchical Model

A Bayesian hierarchical model is defined as a model where the parameters of the prior distribution are also assigned a prior distribution, the so called hyperprior, with parameters that are called hyperparameters (e.g. Ntzoufras, 2009, p. 305-306). They are often used when a certain parameter of interest is estimated in K different units (such as the subgroups in a basket trial) when it is implausible that the unit specific parameters θ_k, $k \in \{1, \ldots, K\}$ are identical or completely unrelated, but it can be assumed that they are exchangeable and arise from a common distribution (e.g. Spiegelhalter et al., 2004, p. 91-92). Prior probabilities are then also assigned to the parameters of this common distribution.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ be the vector of the parameters in the K subgroups and $\boldsymbol{\zeta}$ be the vector of hyperparameters. If the exchangeability assumption holds, the components of $\boldsymbol{\theta}$ are conditionally independent given $\boldsymbol{\zeta}$ due to De Finetti's theorem (e.g. Bernardo, 1996). Unconditionally, however, the specification of a hyperprior leads to an association among the components of $\boldsymbol{\theta}$ (Thall et al., 2003). Specifically, the unconditional prior distribution of $\boldsymbol{\theta}$, $\pi_0(\boldsymbol{\theta})$ is:

$$\pi_0(\boldsymbol{\theta}) = \int \prod_{i=1}^{K} \pi(\theta_i|\boldsymbol{\zeta}) \cdot \pi_0(\boldsymbol{\zeta}) d\boldsymbol{\zeta}.$$

This also leads to dependent posterior distributions:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\int f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\zeta}) d\boldsymbol{\zeta}}{\int f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\zeta}) d\boldsymbol{\theta} d\boldsymbol{\zeta}},$$

where $f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ is the joint density of the data vector $\boldsymbol{y}$ and the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. This dependency results in subgroup-specific estimates that are closer to the overall mean effect than if they were estimated independently, which is called shrinkage. It also results in narrower subgroup-specific credibility intervals (e.g. Lunn et al., 2013, p. 221 - 222).

2.2.3 Bayesian Model Averaging

Model averaging is a Bayesian technique that can be applied when there is a set of plausible models for a parameter θ (e.g. Hoeting et al., 1999). Instead of selecting one of the models, the posterior distribution $\pi(\theta|\boldsymbol{y})$ for θ is computed as a weighted average of several models. Specifically:

$$\pi(\theta|\boldsymbol{y}) = \sum_{j=1}^{J} \pi(\theta|M_j, \boldsymbol{y}) \cdot \pi(M_j|\boldsymbol{y}),$$

where $M_1, \dots M_J$ are the different models, $\pi(\theta|M_j, \boldsymbol{y})$ is the posterior distribution of θ given model M_j and $\pi(M_j|\boldsymbol{y})$ is the posterior distribution of M_j. Hence, the posterior probabilities of the models are used as weights. Applying Bayes' theorem, these are found as:

$$\pi(M_j|\boldsymbol{y}) = \frac{L(M_j|\boldsymbol{y}) \cdot \pi_0(M_j)}{\pi(\boldsymbol{y})} = \frac{L(M_j|\boldsymbol{y}) \cdot \pi_0(M_j)}{\sum_{k=1}^{J} L(M_k|\boldsymbol{y}) \cdot \pi_0(M_k)}, \tag{2.4}$$

where $\pi_0(M_j)$ is the prior probability for model M_j and $L(M_j|\boldsymbol{y})$ is the marginal likelihood of the data given model M_j. This marginal likelihood is found as:

$$L(M_j|\boldsymbol{y}) = \int_0^1 L(M_j|\boldsymbol{y}, \xi_j) \cdot \pi_0(\xi_j|M_j) d\xi_j, \tag{2.5}$$

where ξ_j is the parameter (or vector of parameters) in model M_j, $L(\xi_j|\boldsymbol{y}, M_j)$ is the likelihood of ξ_j given the data and model M_j and $\pi_0(\xi_j|M_j)$ is the prior of ξ_j in model M_j.

2.3 Jensen-Shannon Divergence

The JSD is a measure to quantify the difference between two probability distributions. It is used in Fujikawa's design (see Section 2.4.2) to determine the amount of information that is shared between two baskets. The JSD is defined based on the Kullback-Leibler divergence (KLD) as follows:

$$\text{JSD}(P_k, P_i) = \frac{1}{2}\text{KLD}\left(P_i, \frac{1}{2}(P_i + P_k)\right) + \frac{1}{2}\text{KLD}\left(P_k, \frac{1}{2}(P_i + P_k)\right). \tag{2.6}$$

For two continuous distributions with densities f_k and f_i the KLD is defined as (e.g. Cover and Thomas, 2006, p. 251):

$$\mathrm{KLD}(f_k, f_i) = \int_{\mathbb{R}} f_k(x) \log \frac{f_k(x)}{f_i(x)} dx.$$

Other than the KLD, the JSD is symmetric. The lower bound of the JSD is 0 and is reached if and only if P_k is equal to P_i almost everywhere (Nielsen, 2021). The upper bound is log(2) (Lin, 1991; Nielsen, 2021). Thus, the upper bound depends on the base of the logarithm that is used in the computation of the KLD and is 1 if only if base 2 logarithms are used.

The JSD can also be extended to measure the divergence of more than two distributions. This extension is applied to information sharing in basket trials in Section 3.1.3.3. The extended JSD is defined as:

$$\mathrm{JSD}(\boldsymbol{P}) = \frac{1}{K} \sum_{i=1}^{K} \mathrm{KLD}(P_i, \bar{P}), \tag{2.7}$$

where $\boldsymbol{P} = (P_1, \ldots, P_K)$ is a vector of K probability distributions and $\bar{P} = \frac{1}{K} \sum_{i=1}^{K} P_i$. Note that for $K = 2$ Equation (2.7) and Equation (2.6) coincide. When the KLD is computed using base K logarithms then $0 \leqslant \mathrm{JSD}(\boldsymbol{P}) \leqslant 1$ (Nielsen, 2021).

2.4 Basket Trial Designs

In this section at first some general considerations for designing basket trials are discussed, before describing the specific designs that will be used in the thesis.

2.4.1 Components of a Basket Trial

Pohl et al. (2021) identify four components - two mandatory and two optional - of basket trials. The first mandatory and - from a statistical perspective - the key component of a basket trial is information sharing between subgroups. This is usually the methodologically most complex component and the main focus in the basket trial literature. The second mandatory component is the final analysis, commonly the decision about whether the treatment is effective in one or several baskets. The two optional components are interim assessments that allow early stopping for futility and efficacy.

While basket trial designs proposed in the literature usually include all four components, Pohl et al. (2021) discuss that the components of different designs can be combined in various ways. For example, the technique that is used to share information between baskets in one design can be combined with the interim decision rule of another design. This increases the complexity in comparing basket trial designs, as the influence of the different components on the operating characteristics has to be considered. The emphasis of this thesis is on the information sharing component. Therefore in the following sections, this is also the focus in the description and discussion of the different basket trial designs. The interim components are also reported as proposed by the authors to give a complete picture of the designs. The possibility to combine the different components in other ways should be kept in mind.

2.4.2 Fujikawa's Design

In Fujikawa's design (Fujikawa et al., 2020), information is shared based on the similarity of the basket-wise posterior distributions $\pi(p_k|r_k)$ for the response probabilities p_k, $k \in \{1, \ldots, K\}$. At first, each basket is analysed individually using a beta-binomial model (see Section 2.2.1). Hence, given a beta prior distribution with shape parameters $s_{1,k}$ and $s_{2,k}$, the posterior distribution for the response rate p_k in basket k is:

$$\pi(p_k|r_k) = \text{Beta}(s_{1,k} + r_k, s_{2,k} + n_k - r_k).$$

To share information between baskets, a beta posterior distribution is used for p_k where the posterior parameters are calculated as weighted sums of the basket-wise posterior parameters:

$$\pi(p_k|\boldsymbol{r}, \boldsymbol{\omega_k}) = \text{Beta}(\sum_{i=1}^{K} \omega_{k,i} \cdot (s_{1,i} + r_i), \sum_{i=1}^{K} \omega_{k,i} \cdot (s_{2,i} + n_i - r_i)), \tag{2.8}$$

where $\boldsymbol{\omega_k} = (\omega_{k,1}, \ldots, \omega_{k,K})$ are the weights. Hence, the weights determine how much information is shared between the baskets. Specifically, $\omega_{k,i}$ specifies how much information is shared between basket k and basket i. Note that $\omega_{k,k}$ is always 1 for all $k \in \{1, \ldots, K\}$, which simply indicates that the entire information that is observed in basket k is used in its analysis.

The weights are derived from the pairwise JSD between the posterior distributions for the response probabilities of the baskets without information sharing. Specifically:

$$\omega_{k,i} = \begin{cases} (1 - \text{JSD}(\pi(p_k|r_k), \pi(p_i|r_i)))^{\varepsilon} & \text{if } (1 - \text{JSD}(\pi(p_k|r_k), \pi(p_i|r_i)))^{\varepsilon} > \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where ε and τ are tuning parameters. Note that Fujikawa et al. (2020) define the term "weight" in a slightly different manner. In their manuscript, the weights are defined as $1 - \text{JSD}(\pi(p_k|r_k), \pi(p_i|r_i))$. The tuning parameters ε and τ are added later in the definition of the posterior distribution but are not seen as part of the calculation of the weights $\omega_{k,i}$. This was changed here, such that the tuning parameters are included in the calculation of the weights for better interpretability and better comparability with weights derived from other methods. Note that this does not change the posterior distribution in any way, but simply changes what is referred to as a weight.

Fujikawa et al. (2020) use the natural logarithm in the computation of the JSD. Thus, the upper bound of the JSD is strictly smaller than 1 and the lower bound of the weights is strictly greater than 0 (see Section 2.3) and depends on the choice of ε. For $\varepsilon = 2$, which is the value suggested by Fujikawa et al., the lower bound is $(1 - \log_e(2))^2 \approx 0.09$. However, for $\varepsilon = 1$ it is already $1 - \log_e(2) \approx 0.31$, hence at least 31% of the information is always shared between baskets if ε is set to 1. Therefore, in the rest of the thesis base 2 is used to compute the logarithms in the JSD, to be able to explore a wider range of values for ε without having lower bounds for the sharing weights greater than 0. Specifically, Fujikawa et al. state $\varepsilon \geqslant 1$ as a requirement, but $\varepsilon \in (0, 1)$ can also be considered with the base 2 logarithm.

The amount of information that is shared between baskets can be tuned through the two parameters ε and τ. ε controls how the amount of shared information decreases when the difference of the posterior distributions increases. Weights resulting from different choices of ε are shown in Figure 2. If $\tau > 0$ would be used, then for a number of responses that results in a weight equal to or less than τ in the figure, the weight would be set to 0.

Note that using $\log_2$ should have minor influence on the performance as ε can be chosen such that the weights with $\log_2$ are similar to the weights with $\log_e$. Fujikawa et al. (2020) suggest

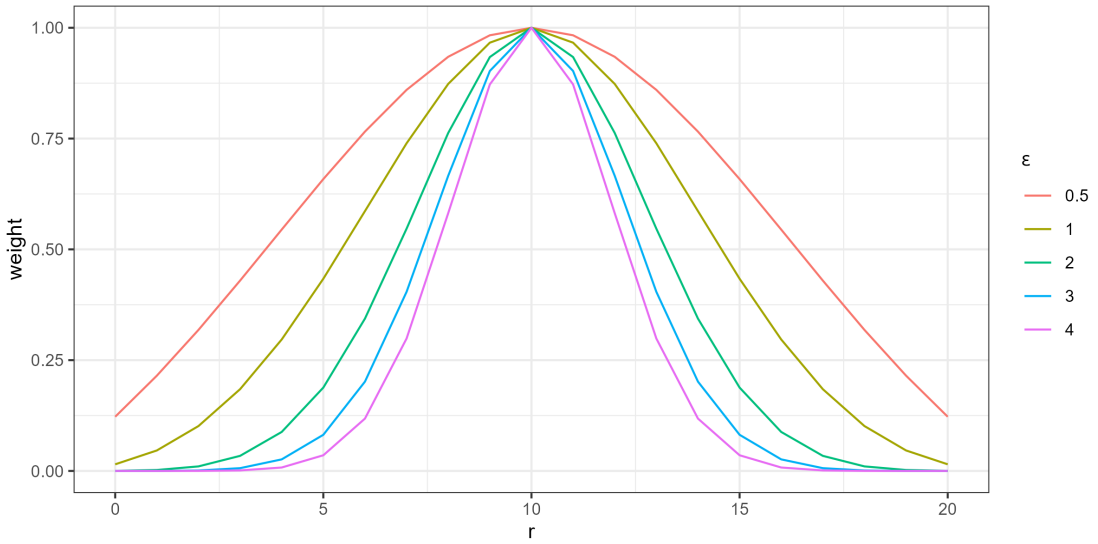

Figure 2: *Weights in Fujikawa's design for different values of ε with $\tau = 0$ and $\log_2$ that are obtained when in one basket the number of responses is 10 and the number of responses r in the other basket varies. A Beta$(1, 1)$ prior and $n = 20$ is used for both baskets.*

$\varepsilon = 2$ and either $\tau = 0$ or $\tau = 0.5$. Figure 3 shows that when $\log_2$ is used together with $\varepsilon = 1.25$, the weights are very similar for most of the range of r. Relevant differences are only seen when the difference in responses is high between the two baskets, as with $\log_2$ the weights are then close to 0 while with $\log_e$ they are not. But this is the intention of using $\log_2$. When $\tau = 0.5$ would be used, this difference would also disappear, as all weights equal to or smaller than 0.5 would be set to 0 then.

The final decision for each basket is based on the posterior distribution with information sharing as given in Equation (2.8):

$$\mathbb{P}(p_k > p_0 | \boldsymbol{r}, \boldsymbol{\omega_k}) \geqslant \lambda,$$

where $\lambda \in (0, 1)$ is a prespecified probability threshold.

Interim analyses in Fujikawa's design are based on the posterior predictive probabilities. The posterior predictive distribution follows (although this was not noted by Fujikawa et al.) a beta-binomial distribution, as described in Section 2.2.1. Specifically, baskets are stopped based on the probability that a result for which the null hypothesis can be rejected will be observed at the end of the trial, given all data observed up until the interim assessment,

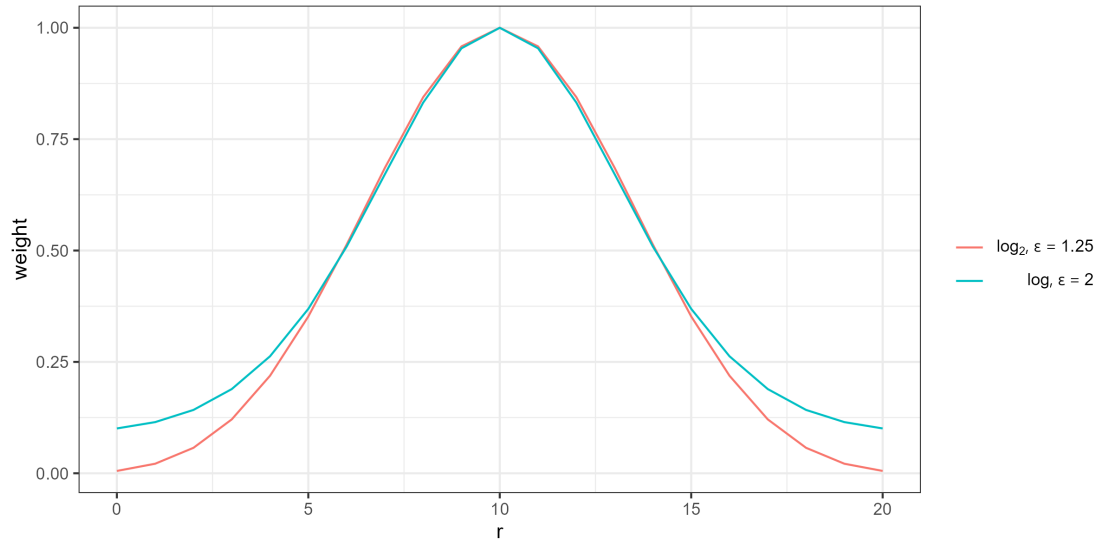

Figure 3: *Weights in Fujikawa's design with* $\tau = 0$ *when* $\varepsilon = 1.25$ *is used with* $\log_2$ *and when* $\varepsilon = 2$ *is used with* $\log_e$ *that are obtained when in one basket the number of responses is 10 and the number of responses* r *in the other basket varies. A Beta*$(1, 1)$ *prior is used for both baskets.*

denoted by $\boldsymbol{r_1} = (r_{1,1}, \ldots, r_{1,K})$. For a trial with one interim analysis, where $n_{1,k}$ denotes the first-stage sample size of basket k, that is:

$$PP_k = \sum_{\tilde{r}_k = r_k - c_k}^{n_k - n_{1,k}} \mathbb{P}(\tilde{r}_k | \boldsymbol{r_1}, \boldsymbol{\omega_{1,k}}),$$

where $\mathbb{P}(\tilde{r}_k | \boldsymbol{r_1}, \boldsymbol{\omega_{1,k}})$ is the posterior predictive probability of $\tilde{r}_k$ given $\boldsymbol{r_1}$ and $\boldsymbol{\omega_{k,1}}$, which is the vector of weights for basket k obtained based on the interim data. Hence, the posterior distribution $\pi(p_k | \boldsymbol{r_1}, \boldsymbol{\omega_{1,k}})$, which shares the information observed in the first stage between baskets, is used in the computation as shown in Equation (2.3). c_k is the critical value for basket k, i.e.:

$$c_k = \min\{r^\star \in \mathbb{N} : \mathbb{P}(p_k > p_0 | r^\star) \geqslant \lambda\}. \tag{2.9}$$

Basket k is then stopped for efficacy if $PP_k > \lambda_e$ and stopped fur futility if $PP_k < \lambda_f$, where $\lambda_e, \lambda_f \in (0, 1)$ are prespecified probability thresholds. Note that baskets which are stopped in an interim assessment are still considered for sharing information in the final analysis.

2.4.3 BMA Design

In the design by Psioda et al. (2021), henceforth referred to as BMA design, all possible cluster configurations of baskets are considered as possible models, and the posterior distributions for the response probabilities are computed as a weighted average of the posterior distributions conditional on the models, as described in Section 2.2.3. Within a cluster, all baskets share a common response probability whereas between clusters the response probabilities differ. Thus, in the case of $K = 3$ baskets for example, there are 5 different models: One in which all baskets share the same response probability, one in which all baskets have different response probabilities and three models where two baskets share the same response probability and the third basket has a distinct response probability.

To derive at first the posterior probabilities of the response probabilities conditional on the models, the following definitions are necessary: Let $p_{(j,l)}$ be l-th distinct response probability in model M_j and $\boldsymbol{p_{(j)}} = (p_{(j,1)}, \ldots, p_{(j,P_j)})$ the vector of all response probabilities in model M_j, where P_j is the number of distinct response probabilities in model M_j. Let further $\Omega_{(j,l)}$ be the subset of the vector $(1, \ldots, K)$ that corresponds to the baskets that share the l-th distinct response rate in model M_j. For example, $\Omega_{(j,1)} = (1, 2)$ and $\Omega_{(j,2)} = 3$ if in model M_j the first and second basket share a response probability and the third basket has a distinct response probability.

The likelihood for $\boldsymbol{p_{(j)}}$ conditional on model M_j can then be written as:

$$L(\boldsymbol{p_{(j)}}|\boldsymbol{r}, M_j) = \prod_{l=1}^{P_j} \left(\prod_{k \in \Omega_{(j,l)}} \binom{n_k}{r_k} \cdot p_{(j,l)}^{r_k} \cdot (1 - p_{(j,l)})^{n_k - r_k} \right).$$

Within each cluster, the data are modelled using a beta-binomial model. Hence, $\pi(p_{(j,p)}|M_j)$, the prior distribution under model M_j, is a beta distribution with shape parameters s_1 and s_2 (that are identical for all models) and thus the posterior distribution for $p_{(j,l)}$ given model M_j is:

$$\pi(p_{(j,l)}|\boldsymbol{r}, M_j) = \text{Beta}(s_1 + \sum_{k \in \Omega_{(j,l)}} r_k, s_2 + \sum_{k \in \Omega_{(j,l)}} (n_k - r_k)),$$

which is also the posterior probability of p_k conditional on model M_j if $k \in \Omega_{(j,l)}$.

To derive the model posterior probabilities, the following prior probabilities are suggested:

$$\pi_0(M_j) \propto \exp(P_j \cdot \psi), \tag{2.10}$$

where ψ is a tuning parameter. Psioda et al. (2021) suggest $\psi \geqslant 0$. With a value of $\psi = 0$ all models have the same prior probability. $\psi > 0$ results in higher prior probabilities for models with more clusters, i.e. less information is shared. However, $\psi < 0$ can also be used, which gives higher prior probabilities to models with less clusters and thus results in more information sharing. Note that in the manuscript of Psioda et al. (2021) there is a typo in the equation for the prior. All results of their paper are, however, based on Equation (2.10) (Psioda, 2023).

Finally, the marginal likelihood (see Equation (2.5)) is:

$$\mathbb{P}(\boldsymbol{r}|M_j) = \prod_{k=1}^{K} \binom{n_k}{r_k} \cdot \prod_{l=1}^{P_j} \frac{\mathcal{B}(s_1 + \sum_{k\in\Omega_{(j,l)}} r_k, s_2 + \sum_{k\in\Omega_{(j,l)}} (n_k - r_k))}{\mathcal{B}(s_1, s_2)}.$$

Hence, all terms from Equation (2.4) have closed-form solutions and thus the model posterior probabilities and therefore also the posterior distributions for all response probabilities can be computed analytically. In the BMA design, a basket is declared active if $\mathbb{P}(p_k > p_0|\boldsymbol{r}) > \lambda$. This decision rule for the final analysis is also used for early stopping for efficacy if one or several interim analyses are conducted. A basket is stopped for futility in an interim analysis, if $\mathbb{P}(p_k > \frac{p_A+p_0}{2}|\boldsymbol{r}) \leqslant \lambda_f$, where p_A is a plausible response probability for an active treatment.

While the posterior quantiles can be calculated analytically, the number of models and thus the computational burden increase exponentially in the number of baskets. The number of models J is:

$$J = \sum_{l=1}^{K} \left(\frac{1}{l!} \sum_{j=0}^{l} (-1)^{l-j} \cdot j^K \binom{l}{j} \right).$$

For 5 baskets, for example, there are 52 models to consider and with 8 baskets the number of models is already 4140. With the R package bmabasket (Alt, 2022), however, posterior quantiles can still be calculated very fast for a moderate number of baskets.

2.4.4 BHM Design

In the design by Berry et al. (2013), in the following referred to as BHM design, Bayesian hierarchical modelling (see Section 2.2.2) is applied. The response probabilities are at first transformed to the logit scale:

$$\theta_k^{\text{BHM}} = \log\left(\frac{p_k}{1-p_k}\right) - \log\left(\frac{p_{A,k}}{1-p_{A,k}}\right). \tag{2.11}$$

$p_{A,k}$ are basket specific target response probabilities, which are the response probabilities assumed under the alternative. Hence the parameters θ_k^{BHM} are defined as the log-odds differences from the target probabilities. This facilitates the comparison of values between baskets when the target probabilities are different and also enables modelling the parameters of the baskets together. The parameters θ_k^{BHM}, $k \in \{1, \ldots, K\}$ are assumed to follow a common normal distribution, i.e.:

$$\theta_k^{\text{BHM}} | \mu, \sigma^2 \sim N(\mu, \sigma^2).$$

As the model is hierarchical, the mean μ and variance σ^2 are also random variables and prior distributions are assigned to them. For μ, a normal prior distribution is specified. Berry et al. (2013) use a mean value close to the null hypothesis and a variance of 100 for the prior, with the justification that this results in an almost non-informative prior distribution. For the variance, different prior distributions can be used. Berry et al. use an inverse-gamma prior for σ^2, but half-normal and half-t for σ or uniform priors for σ^2 or σ are also possible (Cunanan et al., 2017; Neuenschwander et al., 2016).

As explained in Section 2.2.2, by modelling all $\theta_k^{\text{BHM}}, k \in \{1, \ldots, K\}$ in a hierarchical model, information is shared between the baskets by shrinking the basket-wise effects to the overall mean effect. How much information is shared depends on the heterogeneity of the results in the baskets and on the prior distribution for σ^2. The larger σ^2 the less information sharing occurs. Ideally, the amount of borrowing would mainly be determined by the observed data with minimal influence of the selected prior. However, in basket trials the number of subgroups is usually not large enough and thus there is a high sensitivity to the prior

parameters. Freidlin and Korn (2013) find that even in a setting with 10 subgroups prior influence is large. Therefore, careful selection of the prior parameters is essential.

The decision rule for the final analysis to determine whether a basket is active is $\mathbb{P}(p_k > p_0|\boldsymbol{r}) > \lambda$. In interim analyses, early stopping for efficacy and futility is decided based on the probability $\mathbb{P}(p_k > \frac{p_0+p_{A,k}}{2}|\boldsymbol{r})$, with probability thresholds λ_e and λ_f for efficacy and futility, respectively.

Other than for the designs based on the beta-binomial model, the posterior distributions of the BHM design have no closed-form solutions (Thall et al., 2003). Therefore, inference is usually conducted using MCMC sampling. The BHM design is implemented in the R package bhmbasket (Wojciekowski, 2022).

2.4.5 EXNEX Design

In the EXNEX (exchangeability-nonexchangeability) design proposed by Neuenschwander et al. (2016), the probabilities are also transformed to the logit scale, though without considering different target response probabilities:

$$\theta_j^{\mathrm{EX}} = \log\left(\frac{p_k}{1-p_k}\right).$$

These parameters are then modelled as a mixture of two distributions, which are weighted by (possibly basket-specific) fixed weights w_k and $1 - w_k$. The first distribution is an exchangeability distribution, which is defined as in the BHM design:

$$\theta_k^{\mathrm{EX}}|\mu, \sigma^2 \sim N(\mu, \sigma^2).$$

A weakly-informative normal prior for μ that is centred at a plausible value is suggested. The variance is chosen such that the marginal variance for θ_k corresponds to approximately one observation (Neuenschwander et al., 2016, online appendix). For σ^2, a half-normal prior is used.

The second distribution, the nonexchangeability distribution, is:

$$\theta_k^{\mathrm{NEX}} \sim N(\mu_k, \sigma_k^2),$$

where each expected value μ_k is also chosen to correspond to a plausible value and each variance σ_k^2 to a value that corresponds to approximately one observation.

The exchangeability distribution is given weight w_k and the nonexchangebaility distribution is given weight $1 - w_k$. By giving more weight to the nonexchangeable part, less information is borrowed. Thus, the amount of borrowing can be tuned more flexibly than in the BHM design, as in the EXNEX design it depends on both w_k and the prior for σ^2. Note that although the weights are fixed, the posterior distribution is a mixture where the weights are updated (e.g. Bolstad, 2007, p. 319 - 321).

No specific decision criteria for the final analysis or the interim assessments are proposed as part of the design. As mentioned in Section 2.4.1, the information sharing component of the EXNEX design can, however, be combined with the final analysis or interim assessment components of any of the other designs discussed in this section. In the following, $\mathbb{P}(p_k > p_0|\boldsymbol{r}) > \lambda$ will be used as the criterion for the final analysis.

As for the BHM design, inference is based on MCMC sampling as there are no closed-form solutions for the posterior. The EXNEX design is also implemented in the bhmbasket R package.

2.5 Operating Characteristics

Although all designs in the previous section are Bayesian or have some Bayesian elements, basket trial designs are usually compared in terms of their frequentist operating characteristics. Thus, it is assumed the responses in each basket occur with a true fixed but unknown response probability. In this section, the operating characteristics used in the thesis to compare the designs are defined and discussed.

2.5.1 Type 1 Error Rate

The type 1 error rate (TOER) is the probability to falsely reject a null hypothesis. In the context of basket trials there are two types of TOERs - the basket-wise TOERs and the family wise TOER (FWER). Considering the hypotheses defined in Equation (2.1), the basket-wise TOER is the probability to wrongly declare an inactive basket as active. The FWER is the

probability to falsely reject at least one out of K null hypotheses, i.e. to incorrectly declare at least one inactive basket as active.

It is important to note that the basket-wise TOERs also depend on the outcomes of all baskets, due to the information sharing. Under the global null hypothesis, i.e. in a scenario where the true response probability is p_0 in all baskets, where a null hypothesis is rejected if $\mathbb{P}(p_k > p_0|i_1, \ldots, i_K) \geqslant \lambda$, where $i_1, \ldots r_K$ are the responses, the TOER for basket k in a single-stage design is:

$$\text{TOER}_k = \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{1}\left(\mathbb{P}(p_k > p_0|i_1, \ldots, i_K) \geqslant \lambda\right) \cdot \mathbb{P}(i_1|p_0) \cdot \ldots \cdot \mathbb{P}(i_K|p_0),$$

where $\mathbb{1}$ is the indicator function and $\mathbb{P}(i|p_0)$ is the probability of a binomial distribution as given in Equation (2.2). Note that $\mathbb{P}(p_k > p_0|i_1, \ldots, i_K) \geqslant \lambda$ is used inside the indicator function, which is the final decision rule of Fujikawa's design. If a different decision rule is used, then the equation has to be adapted accordingly. Note further that design specific parameters necessary for the computation of the posterior probabilities, such as the weights $\omega_{k,i}$ in Fujikawa's design, are omitted in the notation here. This concerns all equations in this section.

The basket-wise TOERs under scenarios where some baskets are active are found by using the alternative response probability $p_{1,k}$ instead of p_0 for the respective baskets in the calculation of the binomial probabilities.

The FWER under the global null hypothesis is:

$$\begin{aligned}\text{FWER} = \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{1}\left((\mathbb{P}(p_1 > p_0|i_1, \ldots, i_K) \geqslant \lambda) \vee \cdots \vee (\mathbb{P}(p_K > p_0|i_1, \ldots, i_K) \geqslant \lambda)\right) \\ \cdot \mathbb{P}(i_1|p_0) \cdot \ldots \cdot \mathbb{P}(i_K|p_0).\end{aligned}$$

To compute the FWER under scenarios where some of the baskets are active, inside the indicator function the probabilities that correspond to the active baskets have to be removed and, as mentioned above, the binomial probabilities have to be changed accordingly.

To protect the FWER, λ can be tuned such that the FWER under the global null hypothesis does not exceed a certain level (also called weak FWER control). In confirmatory trials, strong FWER control, which in the context of basket trial means controlling the FWER also for any scenario where some of the baskets are active, is required. However, there are currently no solutions for strong FWER control in basket trials where information is shared based on observed similarity. Since basket trials are usually exploratory trials, strong FWER is commonly not desired. If a stronger FWER control than just under the global null hypothesis is wanted, then a weighted mean of FWERs under different scenarios can be calculated and controlled (Kaizer et al., 2021). In this thesis, FWER always refers to the FWER under the global null hypothesis.

2.5.2 Power

The power is the probability to correctly reject a null hypothesis, i.e. considering the hypotheses in Equation (2.1), to correctly declare a basket as active. While experiment-wise power can also be considered, only basket-wise power is discussed and used in the following. As the TOER, the basket-wise power also depends on the outcomes of all other baskets as a result of the information sharing. Under a global alternative hypothesis, where the true response probability in all baskets is p_1, the power for basket k in a single-stage trial is:

$$\text{Pow}_k = \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{1}\left(\mathbb{P}(p_k > p_0 | i_1, \ldots, i_K) \geqslant \lambda\right) \cdot \mathbb{P}(i_1|p_1) \cdot \ldots \cdot \mathbb{P}(i_K|p_1).$$

Power can also be calculated under scenarios where some of the baskets are inactive or all baskets are active, but with different response probabilities, by replacing p_1 in the calculation of the binomial probabilities with the respective probabilities.

Power is a standard operating characteristic to compare different designs e.g. in two-arm trials. In basket trials, however, looking at the power alone gives an incomplete picture. This is because, as mentioned in the previous section, the FWER can usually only be controlled in the weak sense and therefore in scenarios with some active and some inactive baskets, higher power in the active baskets in one design is often accompanied by higher type 1 error inflation in the inactive baskets. Thus, when comparing power, TOER must always be taken into consideration.

2.5.3 Expected Number of Correct Decisions

A less frequently used operating characteristic is the expected number of correct decisions (ECD) (e.g. Broglio et al., 2022). If a null hypothesis is true, then a "correct decision" means that it is not rejected and if the null hypothesis is false, then a "correct decision" means that it is rejected. When K hypotheses are tested, there are at most K correct decisions to be made and thus ECD is a value between 0 and K. ECD is a useful operating characteristic for basket trials, as it in a sense combines TOER and power and thus avoids the problem with comparing power between designs mentioned in the last section. However, as it combines the results of all baskets, ECD is always an operating characteristic for the whole trial.

Let $p_{1,1}, \ldots, p_{1,K}$ be the true response probabilities in the K baskets. Assume, without loss of generality, that $p_{1,1} = p_0$ and $p_{1,k} > p_0$ for $2 \leqslant k \leqslant K$, hence the first basket is inactive and all other baskets are active. Then, for a single-stage trial, the ECD is computed as:

$$\begin{aligned}\text{ECD} = \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \Big(& \mathbb{1}(\mathbb{P}(p_1 > p_0 | i_1, \ldots, i_K) < \lambda) + \mathbb{1}(\mathbb{P}(p_2 > p_0 | i_1, \ldots, i_K) \geqslant \lambda) + \\ & \ldots + \mathbb{1}(\mathbb{P}(p_K > p_0 | i_1, \ldots, i_K) \geqslant \lambda) \Big) \cdot \mathbb{P}(i_1 | p_{1,1}) \cdot \ldots \cdot \mathbb{P}(i_K | p_{1,K}).\end{aligned}$$

Note that the term inside the first indicator function is $\mathbb{P}(p_1 > p_0 | i_1, \ldots, i_K) < \lambda$ since it is assumed that this basket is inactive, and thus a non-rejection of the null hypothesis is a correct decision. The equation has to be adapted according to the assumed response rate scenario.

2.6 Power Prior

2.6.1 General Formulation for a Single and Multiple Historical Studies

The power prior method was developed as a way to generate informative priors based on historical data (Ibrahim and Chen, 2000). The approach assumes that the true parameter values of the historical data and the new data are close, so that the historical data are informative about the parameter of interest θ (Gravestock and Held, 2017). The power prior can, among other things, be applied to generalised linear models, mixed models and Cox proportional hazard models (Ibrahim et al., 2015). To build a power prior for a parameter

of interest θ, an initial prior $\pi_0(\theta)$ and the likelihood $L(\theta|\boldsymbol{y_0})$ of the historical data $\boldsymbol{y_0} = (y_{0,1}, \ldots, y_{0,n_0})$, with n_0 being the sample size of the historical data set, are required. The power prior is then defined as:

$$\pi(\theta|\boldsymbol{y_0}, \omega) \propto L(\theta|\boldsymbol{y_0})^{\omega} \cdot \pi_0(\theta),$$

where $\omega \in [0, 1]$ is a parameter that determines how much weight is given to the historical data. If $\omega = 0$, then the power prior reduces to the initial prior and therefore no information is shared. If $\omega = 1$, the historical data are used completely, hence the power prior is then the posterior distribution of the historical data given the initial prior $\pi_0(\theta)$. When the data arise from a distribution from the exponential family, ω is the percentage of information that is used from the historical data (Pan et al., 2017).

The weight parameter ω can either be treated as a random parameter or calculated based on the current and the historical data and then treated as a fixed value. Neuenschwander et al. (2009) found that when ω is random, information sharing is only moderate even when historical and current data are in complete agreement and sample sizes are large and therefore do not recommend this approach. In the following, only the case with fixed ω will be discussed. When ω is a fixed value, computation of the posterior distribution is much simpler and in some cases, e.g. when data are shared between binomial data sets, power prior and posterior have closed-form solutions as is shown in the next section.

The power prior approach can also be used when several historical data sets are available. Denote the historical data sets by $\boldsymbol{Y_0} = (\boldsymbol{y_{0,1}}, \ldots, \boldsymbol{y_{0,H}})$, where $H \geqslant 2$ is the number of historical data sets. Further let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$ be the respective weight parameters. Then the power prior extends to:

$$\pi(\theta|\boldsymbol{Y_0}, \boldsymbol{\omega}) \propto \prod_{i=1}^{H} L(\theta|\boldsymbol{y_{0,i}})^{\omega_i} \cdot \pi_0(\theta).$$

Hence, each historical data set can be weighted with a different parameter ω_h, where $\omega_h \in [0, 1]$ for all $h \in \{1, \ldots, H\}$.

The use of the power prior is theoretically justified by the fact that it minimises a convex sum of two KLDs including the initial prior and the posterior distribution where the current and the historical data are pooled (Ibrahim et al., 2003). More specifically, let $d_0 = \pi_0(\theta)$ and $d_1 = \pi(\theta|\boldsymbol{y_0}, \omega = 1)$. Let further $d = d(\theta)$ be a density function and ω be non-random. Then the expression:

$$(1 - \omega) \cdot \text{KLD}(d, d_0) + \omega \cdot \text{KLD}(d, d_1),$$

with $\omega \in [0, 1]$, is minimised if d is set to:

$$\pi(\theta|\boldsymbol{y}, \boldsymbol{y_0}, \omega) \propto L(\theta|\boldsymbol{y}) \cdot L(\theta|\boldsymbol{y_0})^{\omega} \cdot \pi_0(\theta),$$

which is the posterior distribution resulting from the power prior, where $\boldsymbol{y} = (y_1, \ldots, y_n)$ refers to the current data here.

A similar result holds for the power prior for multiple historical data sets. Let e_h be a vector of length H where the h-th element is 1 and all other elements are 0. Thus, if $\boldsymbol{\omega} = e_h$, the h-th historical data set is given weight 1 and all other historical data sets are given weight 0. Now let $d_h = \pi(\theta|\boldsymbol{y}, \boldsymbol{Y_0}, \boldsymbol{\omega} = e_h)$. Then the sum:

$$\left(1 - \sum_{i=1}^{H} \omega_i\right) \cdot \text{KLD}(d, d_0) + \sum_{i=1}^{H} \omega_i \cdot \text{KLD}(d, d_i),$$

where $\sum_{i=1}^{H} \omega_i \leqslant 1$ is minimised for d equal to

$$\pi(\theta|\boldsymbol{y}, \boldsymbol{Y_0}, \boldsymbol{\omega}) \propto L(\theta|\boldsymbol{y}) \cdot \left(\prod_{i=1}^{H} L(\theta|\boldsymbol{y_{0,i}})^{\omega_i}\right) \cdot \pi_0(\theta),$$

which is the posterior distribution resulting from the power prior in the case of multiple historical data sets.

2.6.2 Power Prior for Binomial Data

When the data are binomial, the power prior and the resulting posterior distribution have a closed-form solution (e.g. Gravestock and Held, 2017). For the case with a single historical data set, let r_0 be the number of responses in the historical data based on n_0 observations.

The initial prior is chosen as $\pi_0(p) = \text{Beta}(s_1, s_2)$. Then the power prior is:

$$\begin{aligned}
\pi(p|r_0, \omega) &\propto L(p|r_0)^\omega \cdot \pi_0(p) \\
&= \left(\binom{n_0}{r_0} \cdot p^{r_0} \cdot (1-p)^{n_0 - r_0} \right)^\omega \cdot \frac{1}{B(s_1, s_2)} \cdot p^{s_1 - 1} \cdot (1-p)^{s_2 - 1} \\
&\propto \left(p^{r_0} \cdot (1-p)^{n_0 - r_0} \right)^\omega \cdot p^{s_1 - 1} \cdot (1-p)^{s_2 - 1} \\
&= p^{\omega \cdot r_0 + s_1 - 1} \cdot (1-p)^{\omega \cdot (n_0 - r_0) + s_2 - 1} \\
&\propto \text{Beta}(s_1 + \omega \cdot r_0, s_2 + \omega \cdot (n_0 - r_0)).
\end{aligned}$$

Since the power prior is a beta distribution, the posterior is also a beta distribution as shown in Section 2.2.1. Specifically, with r being the number of responses in the current data set based on n observations:

$$\pi(p|r, r_0, \omega) = \text{Beta}(s_1 + r + \omega \cdot r_0, s_2 + (n - r) + \omega \cdot (n_0 - r_0)).$$

Hence, $\omega \cdot 100\%$ of the responses and non-responses observed in the historical data are used in the computation of the posterior when the power prior is applied. The effective sample size (ignoring the information contained in the initial prior) is thus $n + \omega \cdot n_0$.

For the case with multiple historical studies (e.g. Gravestock and Held, 2019) let $\boldsymbol{r_0} = (r_{0,1}, \dots r_{0,H})$ be the H historical data sets (i.e. the vector of responses in the H studies) based on sample sizes $\boldsymbol{n_0} = (n_{0,1}, \dots, n_{0,H})$. Again, the initial prior is a beta distribution. Then:

$$\begin{aligned}
\pi(p|\boldsymbol{r_0}, \boldsymbol{\omega}) &\propto \prod_{i=1}^{H} L(p|r_{0,i})^{\omega_i} \cdot \pi_0(p) \\
&= \prod_{i=1}^{H} \left(\binom{n_{0,i}}{r_{0,i}} \cdot p^{r_{0,i}} \cdot (1-p)^{n_{0,i} - r_{0,i}} \right)^{\omega_i} \cdot \frac{1}{B(s_1, s_2)} \cdot p^{s_1 - 1} \cdot (1-p)^{s_2 - 1} \\
&\propto \prod_{i=1}^{H} \left(p^{r_{0,i}} \cdot (1-p)^{n_{0,i} - r_{0,i}} \right)^{\omega_i} \cdot p^{s_1 - 1} \cdot (1-p)^{s_2 - 1} \\
&= p^{\sum_{i=1}^{H} \omega_i \cdot r_{0,i} + s_1 - 1} \cdot (1-p)^{\sum_{i=1}^{H} \omega_i \cdot (n_{0,i} - r_{0,i}) + s_2 - 1} \\
&\propto \text{Beta}(s_1 + \sum_{i=1}^{H} \omega_i \cdot r_{0,i}, s_2 + \sum_{i=1}^{H} \omega_i \cdot (n_{0,i} - r_{0,i})).
\end{aligned}$$

And the posterior distribution is therefore:

$$\pi(p|r, \boldsymbol{r_0}, \boldsymbol{\omega}) = \text{Beta}(s_1 + r + \sum_{i=1}^{H} \omega_i \cdot r_{0,i}, s_2 + (n - r) + \sum_{i=1}^{H} \omega_i \cdot (n_{0,i} - r_{0,i})).$$

Thus, the effective sample size in this case is $n + \sum_{i=1}^{H} \omega_i \cdot n_{0,i}$.

2.6.3 Calculation of the Weights

A range of different ideas for calculating the weight ω based on the the current and historical data have been suggested (see e.g. Ibrahim et al., 2015; Bennett et al., 2021; Thompson et al., 2021). Most approaches were proposed for the case where only a single historical study is available, but the methods can also be applied to several historical studies by calculating each ω_h separately based on the the h-th historical data set and the current data. In the following, two approaches are introduced. One which was designed for borrowing from a single historical study and can be adapted very flexibly and one with an explicit extension for multiple historical studies. These are later adapted to basket trials in Section 3.1.3.

2.6.3.1 Calibrated Power Prior

Pan et al. (2017) suggested the calibrated power prior (CPP) for borrowing from a single historical study, where ω is calculated based on a prespecified function that quantifies the similarity between the current and the historical data. They discuss the case of normal and binomial data. At first, the Kolmogorov-Smirnov (KS) test statistic between the two data sets is calculated. For that, let $D(t)$ and $D_0(t)$ be the empirical distribution functions of the current and the historical data, respectively. Thus:

$$D(t) = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \leqslant t)}{n},$$
$$D_0(t) = \frac{\sum_{i=1}^{n_0} \mathbb{1}(y_{0,i} \leqslant t)}{n_0}.$$

To calculate the KS test statistic S_{KS}, let $y_{(1)}, \ldots, y_{(n+n_0)}$ be the ordered values of the combined data points of the current and the historical study. Then:

$$S_{\text{KS}} = \max_{i \in \{1,\ldots,n+n_0\}} \{|D(y_{(i)}) - D_0(y_{(i)})|\}.$$

Note that for binomial data the KS statistic is simply the absolute rate difference between the two samples. The KS statistic takes values between 0 and 1 where larger values indicate larger differences. In the next step, the KS statistic is scaled in the following way:

$$S = \max(n, n_0)^{1/4} \cdot S_{\text{KS}}. \tag{2.12}$$

Finally, ω is found as:

$$\omega = \frac{1}{1 + \exp(a + b \cdot \log(S))},$$

where $a \in \mathbb{R}$ and $b > 0$ are tuning parameters. $b > 0$ is required to receive weights that are strictly monotonically decreasing in S_{KS}. For a, values smaller than 0 are possible, but negative values lead to very large weights even when S_{KS} is large.

The scaling step in Equation (2.12) is justified by the fact that with this definition ω converges to 1 if the current and historical data arise from the same underlying distribution and to 0 if they do not, for n and $n_0 \to \infty$.

Pan et al. (2017) propose an algorithm for calibrating the tuning parameters a and b based on the historical data. However, this cannot be applied to basket trials as, in the setting discussed in this thesis, there are no data available in the planning phase. Thus, this is not further discussed here.

2.6.3.2 Weights Based on the Maximum Marginal Likelihood

For a setting in which a single historical study is available, Gravestock and Held (2017) suggested to choose ω as the value that maximises its marginal likelihood. In the general case, that is:

$$\omega = \arg\max_{\omega \in [0,1]} L(\omega | \boldsymbol{y}, \boldsymbol{y_0}),$$

where

$$\begin{aligned} L(\omega | \boldsymbol{y}, \boldsymbol{y_0}) &= \int L(\theta | \boldsymbol{y}) \cdot \pi(\theta | \omega, \boldsymbol{y_0}) d\theta \\ &= \frac{\int L(\theta | \boldsymbol{y_0}) \cdot L(\theta | \boldsymbol{y_0})^{\omega} \pi_0(\theta) d\theta}{\int L(\theta | \boldsymbol{y_0})^{\omega} \cdot \pi_0(\theta) d\theta}. \end{aligned}$$

Constraining ω to the interval $[0, 1]$ is necessary, as otherwise the weight can exceed these limits.

For binomial data:

$$L(\omega|r, r_0) = \int_0^1 L(p|r) \cdot \pi(p|\omega, r_0)dp.$$

Since $L(p|r_0)$ is a binomial likelihood and the posterior $\pi(p|\omega, r_0)$ is a beta distribution, the marginal likelihood $L(\omega|r, r_0)$ follows by definition a beta-binomial distribution (e.g. Bernardo and Smith, 2000, p. 117), with parameters $s_1 + \omega \cdot r_0$, $s_2 + \omega \cdot (n_0 - r_0)$ and n.

Gravestock and Held (2019) investigate how to use the maximum marginal likelihood (MML) approach in a setting with multiple historical studies. They apply the MML weights for a single historical study by estimating ω_h separately for each data set and by pooling all historical data sets and estimate one ω based on this pooled data set. However, they find that these approaches are not ideal and instead suggest an extension of the method. The weights are also found by maximising the marginal likelihood, but of the whole vector of weights $\boldsymbol{\omega}$ given all historical data sets and the current data. Hence:

$$\boldsymbol{\omega} = \arg \max_{\boldsymbol{\omega} \in [0,1]^H} L(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{Y_0}),$$

where

$$L(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{Y_0}) = \int L(\theta|\boldsymbol{y}) \cdot \pi(\theta|\boldsymbol{\omega}, \boldsymbol{Y_0})d\theta.$$

And specifically for binomial data:

$$L(\boldsymbol{\omega}|r, \boldsymbol{r_0}) = \int L(p|r) \cdot \pi(p|\boldsymbol{\omega}, \boldsymbol{r_0})dp.$$

As the posterior $\pi(p|\boldsymbol{\omega}, \boldsymbol{r_0})$ is also a beta distribution, this marginal likelihood also follows a beta-binomial distribution, the parameters are $s_1 + \sum_{i=1}^{H} \omega_i \cdot r_{0,i}$, $s_2 + \sum_{i=1}^{H} \omega_i \cdot (n_{0,i} - r_{0,i})$ and n. Gravestock and Held (2019) argue that by using the marginal likelihood given all historical data sets, the heterogeneity of the historical studies is taken into account and the generated weights combine the historical data such that they have the highest compatibility with the current data.

Chapter 3

Results

In this chapter the results of the thesis are presented. The two main topics are the application of the power prior approach for the analysis of basket trials (Section 3.1) which includes a comprehensive comparison study (Section 3.2) and the investigation of nonmonotonicity as a result of sharing information between baskets (Section 3.3). Two R packages that were developed as part of the thesis are presented in Section 3.4 and Section 3.5.

Note that this chapter also includes new methods and the setup of the comparison study, since these are also results of the thesis. These parts could have also been presented in the previous chapter as methods. However, for the purpose of a continuous narrative and to better distinguish between methods from the literature and methods and ideas developed as part of this thesis, these sections are presented in this chapter.

3.1 Power Priors for Information Sharing in Basket Trials

In this section, the power prior method presented in Section 2.6.1 is adapted such that it facilitates information sharing in basket trials. The connection between Fujikawa's design and the power prior approach is then established. Afterwards, different strategies to compute the weights in the power prior approach, which determine the amount of information that is shared between baskets, are derived.

3.1.1 Adapting the Power Prior Approach to Basket Trials

Power priors, as explained in Section 2.6, are a method for building informative priors based on one or several historical studies that are related to the current study. When in the context of basket trials for a certain basket k the outcomes of all other baskets $j \in \{1, \dots, K\} \backslash k$ are treated in the same way as the historical data sets in the original power prior approach, then many power prior methods can be directly applied to the analysis of basket trials by simply changing the notation.

Let $\boldsymbol{\omega_k} = (\omega_{k,1}, \dots, \omega_{k,K})$ be the vector of weights that determine the amount of information that is shared between basket k and all other baskets. This includes $\omega_{k,k}$ which is always 1. The general formulation of the power prior for basket k is then:

$$\pi(p_k|\boldsymbol{r}_{[-\boldsymbol{k}]}, \boldsymbol{\omega}_{\boldsymbol{k},[-\boldsymbol{k}]}) \propto \left(\prod_{i \in \mathcal{I}_k} L(p_k|r_i)^{\omega_{k,i}} \right) \cdot \pi_0(p_k), \tag{3.1}$$

where the subscript $[-k]$ means that the k-th element of the vector is excluded and $\mathcal{I}_k = \{1, \dots, K\} \backslash k$. The k-th element of $\boldsymbol{r}$ and $\boldsymbol{\omega_k}$ is excluded since they refer to basket k itself, but the data of basket k are not part of the power prior. Note that the notation could be changed, such that $\boldsymbol{\omega_k}$ does not include $\omega_{k,k}$, which seems natural since it is always 1 and could make notation easier since then the subscript $[-k]$ would not be necessary for the weights in the equation above. However, then the j-th element of $\boldsymbol{\omega_k}$ would not correspond to the weight that is shared between basket k and basket j if $j > k$ which would make the notation more complicated in other ways.

The posterior distribution for basket k resulting from the power prior in Equation (3.1) is then (compare with the results from Section 2.6.2):

$$\pi(p_k|\boldsymbol{r}, \boldsymbol{\omega_k}) = \text{Beta}(s_{1,k} + \sum_{i=1}^{K} \omega_{k,i} \cdot r_i, s_{2,k} + \sum_{i=1}^{K} \omega_{k,i} \cdot (n_i - r_i)). \tag{3.2}$$

Hence, the parameters of the posterior distribution of basket k are calculated as a weighted sum of the responses and non-responses of all baskets plus the prior parameters of basket k.

Note that in the following, as in Fujikawa's design, $\mathbb{P}(p_k > p_0|\boldsymbol{r}, \boldsymbol{\omega_k}) \geqslant \lambda$ is used as the criterion to reject the null hypothesis in the final analysis.

3.1.2 Connection Between Fujikawa's Design and Power Priors

The posterior distribution given in Equation (3.2) resulting from applying the power prior and the posterior distribution in Equation (2.8) which is used in Fujikawa's design to share information between baskets are clearly closely related to each other. Fujikawa et al. (2020), however, do not mention this connection to power priors. In both methods, information is generally shared in the same manner by computing a weighted sum of the available information. The difference between the two approaches is, however, that in the posterior distribution of Fujikawa's design, the prior parameters of all baskets are also part of the weighted sums, while with the power prior approach only the prior parameters of basket k are used in the posterior distribution of basket k. Thus, not only the observed data, but also the prior information is shared in Fujikawa's design. Of course weights based on the JSD can also be used for sharing information with power priors to align Fujikawa's design and the power prior approach.

Whether sharing prior information is appropriate may depend on how the prior parameters are chosen. If they are based on trial data from earlier phases, then also sharing this information seems sensible. If, however, no prior information is available and a non-informative beta prior is used, including the prior parameters of the other baskets in the posterior distribution is not reasonable, and may potentially introduce additional bias. For example, Fujikawa et al. (2020) use $\text{Beta}(1, 1)$ priors, which contain as much information as observing two patients, one with a response and one without a response. Thus, this prior indicates a response rate of 0.5. If this prior is used for all baskets in a basket trial with a large number of baskets and small sample sizes per basket, then the influence of sharing the prior information can be non-negligible, especially when the observed response rates are either far below or above 0.5. On the other hand, if a prior distribution with much smaller prior parameters than 1, for example $\text{Beta}(0.001, 0.001)$, is used then the influence of the shared prior information is irrelevant and Fujikawa's design and the power prior approach will lead to almost identical results.

It may seem that the posterior distribution from Fujikawa's design can also be expressed as a posterior distribution resulting from a power prior, by including the prior distributions of the other baskets in the power prior. But this does not lead to the same posterior distribution

as used in Fujikawa's design. The power prior which includes all prior distributions is:

$$
\begin{aligned}
\pi(p_k|\boldsymbol{r}_{[-\boldsymbol{k}]}, \boldsymbol{\omega}_{\boldsymbol{k},[-\boldsymbol{k}]}) &\propto \left(\prod_{i \in \mathcal{I}_k} \left(L(p_k|r_i) \cdot \pi_0^i(p_k) \right)^{\omega_{k,i}} \right) \cdot \pi_0^k(p_k) \\
&= \prod_{i \in \mathcal{I}_k} \left(\binom{n_i}{r_i} \cdot p_k^{r_i} \cdot (1-p_k)^{n_i - r_i} \cdot \frac{1}{B(s_{1,i}, s_{2,i})} \cdot p_k^{s_{1,i}-1} \cdot (1-p_k)^{s_{2,i}-1} \right)^{\omega_{k,i}} \cdot \\
&\quad \frac{1}{B(s_{1,k}, s_{2,k})} \cdot p_k^{s_{1,k}-1} \cdot (1-p_k)^{s_{2,k}-1} \\
&\propto \prod_{i \in \mathcal{I}_k} \left(p_k^{r_i} \cdot (1-p_k)^{n_i - r_i} \cdot p_k^{s_{1,i}-1} \cdot (1-p_k)^{s_{2,i}-1} \right)^{\omega_{k,i}} \cdot p_k^{s_{1,k}-1} \cdot (1-p_k)^{s_{2,k}-1} \\
&= p_k^{s_{1,k}-1+\sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot (r_i + s_{1,i} - 1)} \cdot (1-p_k)^{s_{2,k}-1+\sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot ((n_i - r_i) + s_{2,i} - 1)} \\
&\propto \text{Beta}(s_{1,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot (r_i + s_{1,i} - 1), s_{2,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot ((n_i - r_i) + s_{2,i} - 1)) \\
&= \text{Beta}(s_{1,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot (r_i + s_{1,i}) - \sum_{i \in \mathcal{I}_k} \omega_{k,i}, \\
&\quad s_{2,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot ((n_i - r_i) + s_{2,i}) - \sum_{i \in \mathcal{I}_k} \omega_{k,i}).
\end{aligned}
$$

Note the superscript i in $\pi_0^i(p_k)$ is used here to indicate that the prior parameters chosen for basket i are used, since the prior parameters may in general differ between baskets. Normally, this superscript is not necessary as the subscript of the parameter indicates which prior distribution is used - e.g. $\pi_0(p_k)$ refers to the prior distribution of basket k. Since in the equation above the power prior is built for basket k using the prior distributions of all baskets, this is not clear and thus the superscript is introduced here.

The posterior distribution is then:

$$
\pi(p_k|\boldsymbol{r}, \boldsymbol{\omega}_{\boldsymbol{k}}) = \text{Beta}(\sum_{i=1}^{K} \omega_{k,i} \cdot (s_{1,i} + r_i) - \sum_{i \in \mathcal{I}_k} \omega_{k,i}, \sum_{i=1}^{K} \omega_{k,i} \cdot (s_{2,i} + (n_i - r_i)) - \sum_{i \in \mathcal{I}_k} \omega_{k,i}),
$$

which differs from the posterior distribution of Fujikawa's design by the term $-\sum_{i \in \mathcal{I}_k} \omega_{k,i}$ in both posterior parameters. If historical data are, however, available for any of the baskets, these could of course also be used by including it as a further likelihood term in the power prior. The historical data can be combined with the new data such that both receive the same weight, or additional weights can be introduced to weight the historical and the current data differently.

3.1.3 Calculation of the Weights for Basket Trials

3.1.3.1 Calibrated Power Prior Weights for Basket Trials

The CPP approach described in Section 2.6.3.1 can easily be applied to basket trials. The CPP method was proposed in the context of a single historical study, but as mentioned in Section 2.6.3, power prior methods for single studies can be extended to the multiple studies case by calculating the weight separately for each historical study. By further extension in the setting of basket trials, the weights are found by calculating the weight individually for each pair of baskets (as it is also done in Fujikawa's design). Thus, in a single-stage trial with K baskets there are $\binom{K}{2}$ weights to be calculated. Specifically, by adapting the notation in Equation (2.6.3.1) to the basket trial setting, the weight $\omega_{k,i}$ that determines how much information is shared between basket k and basket i is:

$$\omega_{k,i} = \frac{1}{1 + \exp(a + b \cdot \log(S_{k,i}))},$$

where $S_{k,i} = \max(n_k, n_i)^{1/4} \cdot S_{\mathrm{KS};k,i}$ is the scaled and $S_{\mathrm{KS};k,i}$ the unscaled KS test statistic between basket k and i.

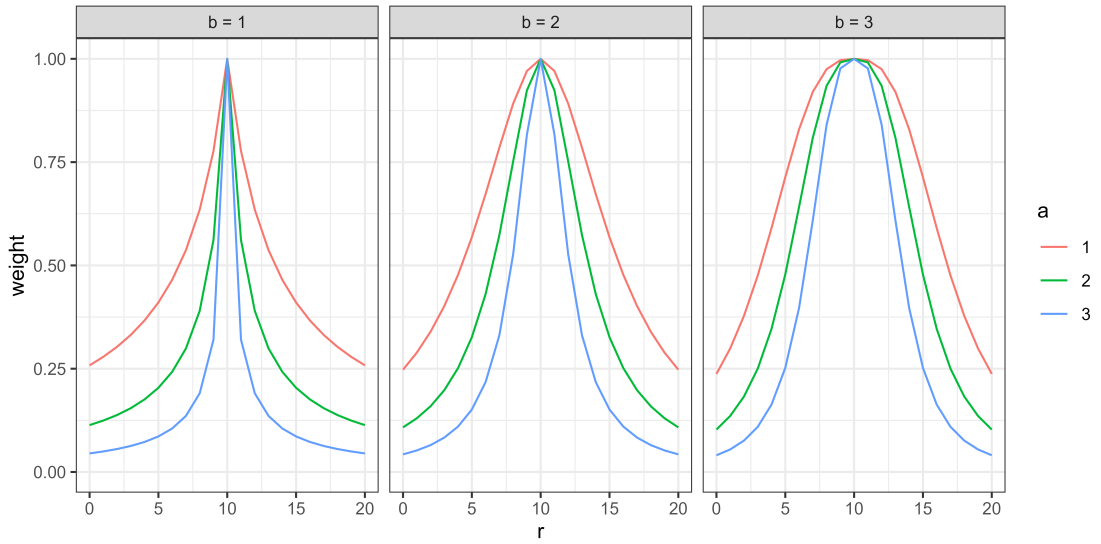

Figure 4: *CPP weights in the power prior design for different values of the tuning parameters a and b that are obtained when in one basket the number of responses is 10 and the number of responses r in the other basket varies. $n = 20$ in each basket.*

Figure 4 displays different weights that are obtained for different choices for a and b.

3.1.3.2 Maximum Marginal Likelihood Weights for Basket Trials

In Section 2.6.3.2, two methods for calculating the power prior weights based on the maximum of the marginal likelihood of ω (for a single historical study) or $\boldsymbol{\omega}$ (for several historical studies) were presented.

The method for a single historical study can again be extended to basket trials by calculating the weights based on pairwise similarity. The weights are then calculated as:

$$\begin{aligned}\omega_{k,i} &= \arg\max_{\omega_{k,i}\in[0,1]} L(\omega_{k,i}|r_k, r_i) \\ &= \arg\max_{\omega_{k,i}\in[0,1]} f_{\text{BetaBin}}(r_k; s_{1,k} + \omega_{k,i}\cdot r_i, s_{2,k} + \omega_{k,i}\cdot(n_i - r_i), n_k)\end{aligned}$$

where $f_{\text{BetaBin}}(\cdot; t_1, t_2, n)$ is the density of a beta-binomial distribution with parameters t_1, t_2 and n.

However, this definition leads to asymmetric weights. For example, consider two baskets each with $n = 20$ and outcomes $r_1 = 9$ and $r_2 = 4$. Given a Beta$(1, 1)$ prior distribution, the resulting weights are $\omega_{1,2} = 0.14$ but $\omega_{2,1} = 0.118$. CPP weights and weights in Fujikawa's design are always symmetric, since they only depend on the KS test statistic and the JSD, respectively, which are both symmetric functions.

In the context of borrowing from a historical study this may not be an issue, since information is only shared in one direction - the current study borrows from the historical study but not the other way round. In basket trials, however, asymmetric weights based on pairwise similarity seem counterintuitive. To receive a symmetric weight function, the weights can be symmetrised by taking the mean of two weights $\omega_{k,i}$ and $\omega_{i,k}$ resulting from the MML approach:

$$\begin{aligned}\omega_{k,i}^{\text{Sym}} = \frac{1}{2}\big(&\arg\max_{\omega_{k,i}\in[0,1]} f_{\text{BetaBin}}(r_k; s_{1,k} + \omega_{k,i}\cdot r_i, s_{2,k} + \omega_{k,i}\cdot(n_i - r_i), n_k) + \\ &\arg\max_{\omega_{k,i}\in[0,1]} f_{\text{BetaBin}}(r_i; s_{1,i} + \omega_{k,i}\cdot r_k, s_{2,i} + \omega_{k,i}\cdot(n_k - r_k), n_i)\big).\end{aligned}$$

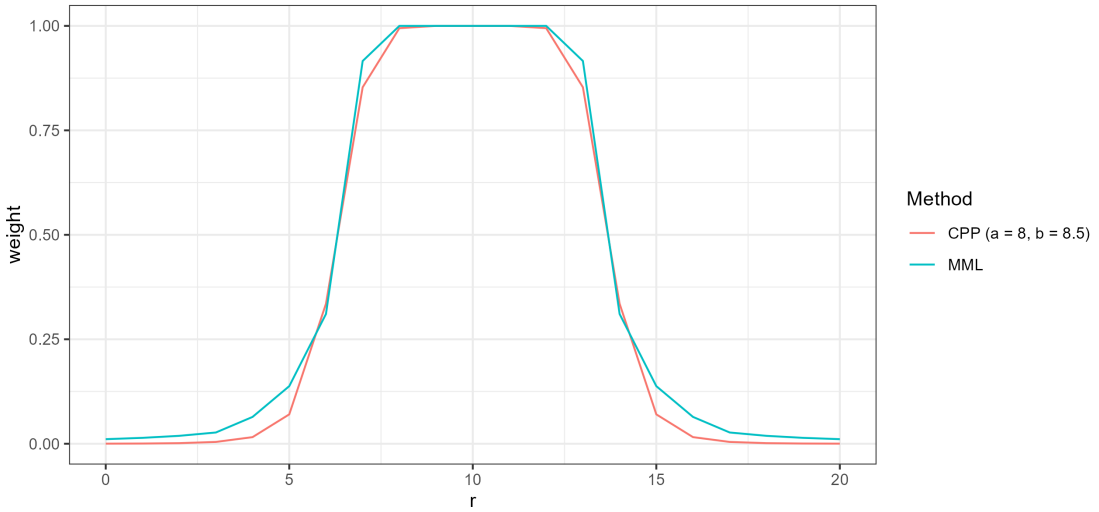

Figure 5: *Symmetrised MML weights and CPP weights with* $a = 8$ *and* $b = 8.5$ *in the power prior design that are obtained when in one basket the number of responses is 10 and the number of responses* r *in the other basket varies. A* $Beta(1, 1)$ *prior and* $n = 20$ *is used for both baskets.*

Note, however, that a very similarly shaped weight function can be obtained by tuning CPP weights accordingly. The symmetrised MML weights and CPP weights with tuning parameters $a = 8$ and $b = 8.5$ are shown in Figure 5.

In the extension of the MML approach for multiple historical studies, the weights are not found based on pairwise similarity. Instead, applying the method to basket trials, for each basket k the weights that determine the amount of information that is shared from all other baskets is found by maximising the marginal likelihood of $\boldsymbol{\omega}_{\boldsymbol{k};[-\boldsymbol{k}]}$ given $\boldsymbol{r}$. Thus:

$$\begin{aligned}\boldsymbol{\omega}_{\boldsymbol{k};[-\boldsymbol{k}]} &= \arg \max_{\boldsymbol{\omega}_{\boldsymbol{k};[-\boldsymbol{k}]} \in [0,1]^{K-1}} L(\boldsymbol{\omega}_{\boldsymbol{k}} | \boldsymbol{r}) \\ &= \arg \max_{\boldsymbol{\omega}_{\boldsymbol{k};[-\boldsymbol{k}]} \in [0,1]^{K-1}} f_{\text{BetaBin}}(r_k; s_{1,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot r_i, s_{2,k} + \sum_{i \in \mathcal{I}_k} \omega_{k,i} \cdot (n_i - r_i), n_k).\end{aligned}$$

Therefore, when this approach is extended to basket trials, the arg max has to be calculated K times, such that every basket is treated as the "current study" once, and all other baskets act as the "historical studies".

When the weights are calculated based on all observations in this way, symmetry is in general not expected. To extend the example above, consider now that four baskets were observed and

the vector of outcomes is $\boldsymbol{r} = (4, 9, 10, 11)$. Now, again given Beta$(1, 1)$ prior distributions, the weight vector for the first basket is $\boldsymbol{\omega_1} = (1, 0.12, 0, 0)$ and the weight vector for the second basket is $\boldsymbol{\omega_2} = (0.71, 1, 1, 1)$. This is, since for finding the weights for basket 1, the marginal likelihood conditions on baskets 2 to 4 which are similar to each other but far away from the results of basket 1. Thus, little information is shared from baskets 2 to 4. However, when calculating the weights for basket 2, the parameters of the beta-binomial density are calculated based on baskets 1, 3, and 4. Between these three baskets there is more heterogeneity, but the results of basket 2 are in between the other three subgroups, thus more information is shared from all other baskets.

3.1.3.3 Weights Based on Overall Heterogeneity for Basket Trials

The power prior design in which weights are based on pairwise similarity between baskets (using either CPP weights or JSD weights as in Fujikawa's design) can be extended to incorporate the finding of Gravestock and Held (2019) in the context of borrowing data from historical studies, that using all available data simultaneously is preferable to only considering the pairwise similarity to derive the weights.

For that purpose, global weights $\omega^\star \in [0, 1]$ can be used, which quantify the overall heterogeneity of the data of all baskets. Larger values for $\omega^\star$ indicate that little heterogeneity is observed between the baskets and a smaller value for $\omega^\star$ means that there is high heterogeneity. The global weight is then combined with the pairwise weights and the information that is shared between two baskets is then found as:

$$\omega'_{k,i} = \omega_{k,i} \cdot \omega^\star.$$

The idea is therefore that less information is shared between baskets when high heterogeneity is observed, even when two baskets have very similar or identical results. Clearly, this still leads to symmetric sharing between baskets, as the same value for $\omega^\star$ is used for the calculation of all weights.

To quantify the overall heterogeneity and calculate $\omega^\star$, one option is to use the extended version of the JSD as given in Equation (2.7). $\omega^\star$ can then be calculated as:

$$\omega^\star = \left(1 - \text{JSD}(\pi(p_1|r_1), \ldots, \pi(p_K|r_K))\right)^{\varepsilon^\star},$$

where $\varepsilon^\star$ is a tuning parameter. A disadvantage of using the JSD to calculate $\omega^\star$ is, however, that it requires numerical integration, which is computationally relatively expensive. When the JSD is only used to compute the pairwise weights, as in Fujikawa's design, this is not an issue since the number of possible weights that can occur is at most $(n+1)^2$ in a single-stage design with equal sample sizes. But since $\omega^\star$ is based on the outcomes of all baskets, the JSD has to be computed for every vector of outcomes. This slows the computation of operating characteristics based on simulation significantly and makes analytical computations infeasible.

An alternative function h that quantifies the overall heterogeneity and is cheaper to compute is derived from the following example and considerations: In a basket trial with $K = 3$ baskets, let $rr_k = r_k/n_k, k \in \{1, \ldots, 3\}$ be the response rates. Let further $rr_{(j)}$ denote the j-th of the ordered response rates and define $d_1 = rr_{(3)} - rr_{(2)}$ and $d_2 = rr_{(2)} - rr_{(1)}$. Hence, d_1 is the difference between the largest response rate and the response rate in the middle and d_2 is the difference between the middle and the smallest response rate, thus $d_1, d_2 \in [0, 1]$. h is defined as a function of d_1 and d_2 and should fulfil the following criteria: First, when there is no heterogeneity, the value of h should be 1. No heterogeneity is present when the response rates of all baskets are identical. Second, when there is maximum heterogeneity, the value of h should be 0. Maximum heterogeneity may be defined in other ways, but one possibility is to define it - in the example with three baskets - as observing response rates of 0, 0.5 and 1, i.e. as an equidistant sequence on the parameter space $[0, 1]$. Thus, mathematically the conditions for h are:

1. $h(0, 0) = 1$,
2. $h(0.5, 0.5) = 0$.

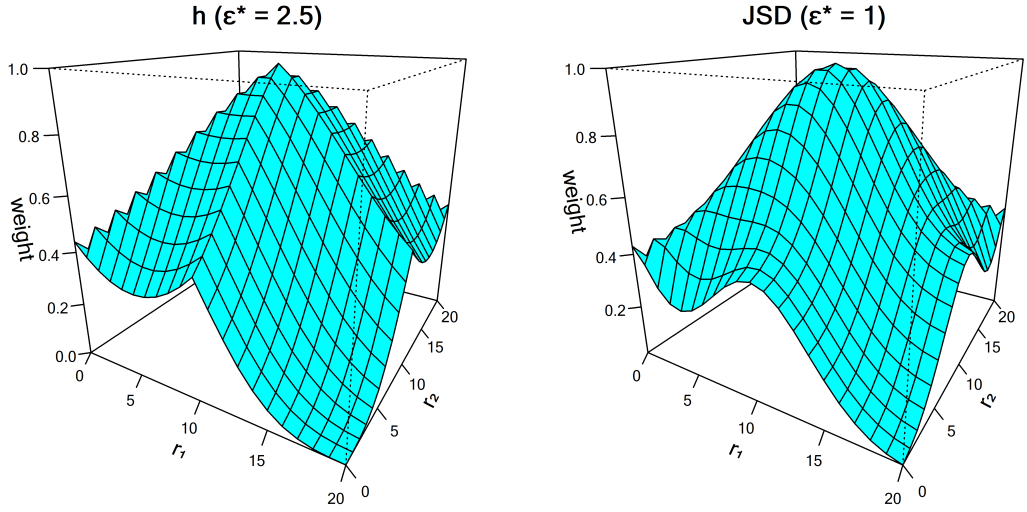

Figure 6: *Global weights based on h with $\varepsilon^\star = 2.5$ and the extended JSD with $\varepsilon^\star = 1$ with $K = 3$ baskets that are obtained when in one basket the number of responses is 10 and the responses r_1 and r_2 vary. A Beta$(1, 1)$ prior and $n = 20$ are used for all baskets.*

The following definition of h satisfies these conditions and also their generalisation to the case of $K > 3$:

$$h(\boldsymbol{d}) = \left(1 - \sum_{i=1}^{K-1} d_i \cdot 10^{\sum_{i=1}^{K-1}(d_i - 1/(K-1))^2}\right)^{\varepsilon^\star}, \tag{3.3}$$

where $\boldsymbol{d} = (d_1, \ldots d_{K-1})$ and $d_i, i \in \{1, \ldots, K-1\}$ are defined analogously to the example with three baskets and $\varepsilon^\star$ is a tuning parameter. The function could also be tuned in other ways. For example, the "10" could be changed to another positive number and the function would still satisfy the two conditions. But this is not further explored to not introduce too many different tuning parameters. Clearly, since h only involves basic arithmetic operations, it is much cheaper to compute than the JSD. Analytical computation of operating characteristics is still feasible when global weights based on h are used. Figure 6 visualises the two global weight functions. When for h the tuning parameter $\varepsilon^\star = 2.5$ and for the JSD it is set to $\varepsilon^\star = 1$, the two functions look relatively similar.

Another option is to set $\omega^\star$ to a fixed value between 0 and 1 without considering the observed heterogeneity, such that more weight is given to the data observed in the basket itself and less information is shared between baskets. This idea is in the spirit of the EXNEX design as described in Section 2.4.5. However, other than the weight w in the EXNEX design, the fixed $\omega^\star$ is not updated based on the observed similarity.

The three options to choose $\omega^\star$ described above (based on the JSD, the heterogeneity function h or as a fixed value) can be combined with any approach to determine pairwise weights. Two natural combinations are to combine $\omega^\star$ based on the JSD with pairwise weights also based on the JSD and to combine $\omega^\star$ based on the heterogeneity function h with pairwise weights based on the CPP, since both h and the CPP weights are based on the response rate differences between baskets.

3.2 Comparison Study

In this section, several variations of the power prior design proposed in the previous section are compared with Fujikawa's design and the EXNEX, BHM and BMA designs.

3.2.1 Setup

Designing a comparison study for different basket trial designs is complex, as the number of design elements as well as the number of prior and tuning parameters involved in the different designs is large. Relevant design elements and parameters include the number and sample sizes of the baskets. If interim assessments should also be considered, then the number, timing and type of interim assessments must be chosen. Furthermore, there is an enormous amount of possible response probability scenarios. In some simulation studies, a single null and a single alternative response probability are set, resulting in $K + 1$ different scenarios, with 0 to K active baskets. However, in practice baskets which respond to the treatment may show different response rates, hence mixed alternative scenarios can also be considered.

A further aspect to consider is TOER control. Often strong control of the FWER, i.e. under all possible configurations of true null and alternative hypotheses, is desired. However, in basket trials with information sharing only weak FWER control, i.e. under the global null hypothesis is feasible as discussed in Section 2.5.1.

Finally, competing designs have to be chosen carefully. In their review, Pohl et al. (2021) discuss around 20 different designs and new designs have been proposed since the article was published. The number of comparison studies is, however, low and little is known about which methods perform best under which circumstances. Therefore, it is difficult to choose

Table 2: *Response probability scenarios considered in the comparison study*

Scenario	Basket 1	Basket 2	Basket 3	Basket 4
Global Null	0.15	0.15	0.15	0.15
Global Alternative	0.4	0.4	0.4	0.4
One in the Middle	0.4	0.4	0.3	0.5
Linear	0.15	0.25	0.35	0.45
Good Nugget	0.15	0.15	0.15	0.4
Bad Nugget	0.15	0.4	0.4	0.4
Half	0.15	0.15	0.4	0.4

designs for a fair comparison. A further problem is that many designs are not implemented in R packages or other software.

The focus of this comparison study is on the different sharing techniques, therefore a single-stage trial without any interim analyses is considered. Equal sample sizes of $n = 20$ were chosen and the number of baskets is set to $K = 4$. The response probability scenarios are inspired by scenarios used in the simulation studies in Berry et al. (2013) and Broglio et al. (2022). The investigated scenarios are shown in Table 2. The null response rate is set to $p_0 = 0.15$. For the Global Alternative scenario the response probability was set to 0.4. The Good Nugget, Half and Bad Nugget scenario represent cases where 1, 2 and 3 baskets are active, respectively, where the active baskets also have a response probability of 0.4. Additionally, a Linear scenario and a One in the Middle scenario were considered. In the former, 3 of the 4 baskets are active with linearly increasing response probabilities. In the One in the Middle scenario, all baskets are active but with different response probabilities.

The following methods were investigated (ordered alphabetically):

- BHM design
- BMA design
- CPP: Power prior design with pairwise weights only, based on the CPP approach,
- CPP-Global: Power prior design with pairwise weights based on the CPP approach, global weights based on h,

- CPP-Nex: Power prior design with pairwise weights based on the CPP approach, fixed global weight,
- EXNEX design
- Fujikawa's design
- JSD-Global: Power prior design with pairwise and global weights, both based on the JSD,
- MML: Power prior design with symmetrised pairwise weights based on the MML approach,
- MML-Global: Power prior design with weights based on the MML approach for multiple historical studies.

The BMA, BHM and EXNEX designs were selected as comparators, since R packages that allow applying these designs are available on the Comprehensive R Archive Network (CRAN). The package bhmbasket (Wojciekowski, 2022) implements the BHM and the EXNEX design. bmabasket (Alt, 2022) implements the BMA design. Fujikawa's design is implemented in the baskexact package (see Section 3.4), which also implements the power prior design CPP, CPP-Global, CPP-Nex and MML weights. The remaining power prior variations are implemented in the basksim package which is described in Section 3.5.

The tuning and prior parameter values were selected following an approach also used in a simulation study by Broglio et al. (2022) using the ECD (see Section 2.5.3) as the target for optimisation. The following steps were applied for each method to select the optimal tuning and prior parameter values:

1. Define a set of potential values for each tuning and prior parameter that should be optimised.
2. For each combination of values, find the smallest posterior probability threshold λ (up to a margin of 0.001) such that the FWER under the global null hypothesis does not exceed $\alpha = 0.05$ (one-sided).

3. For each combination of values and the respective λ found in step 2, compute the ECD under each scenario. Calculate the mean ECD across all scenarios.

4. Select the combination of tuning and prior parameter values that results in the highest mean ECD.

Based on the tuning and prior parameter values found optimal in this procedure, the designs were compared based on the ECDs, rejection rates and posterior means.

Operating characteristics for the power prior design with CPP, CPP-Global, CPP-Nex and MML weights as well as Fujikawa's design were calculated analytically. For all other methods, simulations were used. For that purpose, 10,000 simulated data sets were created for each scenario, which were used for all methods. For the BMA design and the power prior design with JSD-Global and MML-Global weights, posterior probabilities were calculated analytically, but for the BHM and EXNEX design posterior probabilities are based on 10,000 MCMC samples plus 1000 discarded burn-in MCMC samples.

For computation of the BHM and EXNEX design, minor modifications of the bhmbasket package were made (modified version available on `https://github.com/lbau7/bhmbasket`). First, in bhmbasket the results are computed using a nested for-loop with parallelisation using the doRNG package (Gaujoux, 2023). However, nested parallel for-loops are not supported according to the package's vignette and thus unwanted correlations between MCMC samples may be introduced. Therefore, the inner loop was changed to a normal for-loop without parallelisation. Second, the number of burn-in samples and the number of chains used in the MCMC sampling in bhmbasket cannot be changed by the user. In the modified version, the number of chains was set to 1 and the number of burn-in samples was set to 1000.

3.2.2 Potential Tuning Parameter Values

Selection of potential tuning and prior parameter values was also guided by Broglio et al. (2022). To limit the number of different parameter value combinations, parameters that have little impact on information sharing were not considered in the optimisation. For all relevant parameters a grid of values was specified.

For all power prior design variations, the prior distributions were set to Beta$(1, 1)$. With CPP, CPP-Global and CPP-Nex weights, the information sharing only depends on the rate differences and therefore the choice of beta prior parameters does not influence the amount of shared information. With JSD-Global weights, the prior has minimal influence on information sharing, since the JSD weights are derived from the individual posterior distributions which depend on the prior choice. However, the numerical integration with R's default integrate function, which is used to compute the JSD, does not converge for some response vectors when $s_1, s_2 < 1$ are used as prior parameters of the beta distribution.

For the pairwise CPP weights, the tuning parameters that influence information sharing are a and b. For each of these two parameters, 6 equidistant values between 0.5 and 3 were tested. This range for the tuning parameter values was determined by informal calculations with $K = 3$ baskets. The same set of values was investigated for the parameter $\varepsilon^\star$ which is used to compute the global heterogeneity component of the CPP-Global weights. For the fixed global weight component of the CPP-Nex weights, 9 equidistant values between 0.1 and 0.9 were considered. This results in 36 different parameter value combinations for CPP weights, 216 different combinations for CPP-Global weights and 288 possible values for CPP-Nex weights. For the JSD-Global weights, information sharing is guided by ε and τ in the computation of the pairwise weights and by $\varepsilon^\star$ in the computation of the global heterogeneity component. For ε and $\varepsilon^\star$, values between 0.5 and 3 in steps of 0.5 were considered, and for τ 6 equidistant values between 0 and 0.5 were used. Hence, 216 different value combinations were tested. In the computation of the MML and MML-Global weights, no tuning parameters are involved. Note that MML weights were used despite their similarity to CPP weights shown in Section 3.1.3.2, but the investigated range of CPP parameter values does not include the values that result in a shape of the weight function similar to that of the MML weights. The choice of the beta prior parameters has some influence on the weights, but to be consistent with the other methods, s_1 and s_2 were held constant at 1. Gravestock and Held (2017) and Gravestock and Held (2019) also used a Beta$(1, 1)$ prior distribution.

For Fujikawa's design, the prior distributions were also set to Beta$(1, 1)$ due to the computational issue mentioned above and since this is also the prior used in Fujikawa et al. (2020).

The values for τ and ε where chosen as for the power prior design with JSD-Global weights, i.e. 36 different combinations were used.

The BMA design also involves beta priors for each cluster. For that, again Beta$(1, 1)$ priors were used. ψ influences the amount of information sharing since it determines the prior probabilities for the models considered in the model averaging. For this parameter 17 values between -4 and 4 with a step size of 0.5 were considered.

The BHM design requires specification of two prior distributions, one for μ and one for σ^2. The distribution for μ was set to $N(-1.3291, 100^2)$. The expected value -1.3921 of this normal distribution is equal to the null hypothesis given $p_0 = 0.15$ and $p_{\text{targ}} = 0.4$ - see Equation (2.11). The variance of 100^2 results in a practically non-informative prior. For σ, a half-normal prior with varying scale parameters ϕ was used. The half-normal prior was used as this is the only prior option implemented in the bhmbasket package, although Berry et al. (2013) use an inverse-gamma prior for σ^2. Broglio et al. (2022) however, also use a half-normal prior for σ. For the scale parameter - again as used by Broglio et al. (2022) - 8 equidistant values ranging from 0.125 to 2 were investigated.

The EXNEX design also requires specification of a distribution for μ and σ^2 and additionally requires K prior distributions, one for the transformed individual response rate θ_k^{NEX} of each basket. For μ and all θ_k^{NEX}, the prior distributions were set to $N(-1.7346, 100^2)$. The mean of this prior distribution is not the same as in the prior used for μ in the BHM design, as the target response probabilities that are used in the BHM design are not used to adjust the transformed response probabilities in the EXNEX design. As a prior for σ, a half-normal prior was used and the same 8 values for ϕ as in the BHM design were considered. For the weight parameter w, 9 equidistant values between 0.1 and 0.9 were taken into consideration. Thus in total 72 parameter value combinations were investigated.

3.2.3 Results of the Comparison Study

The prior and tuning parameter values that resulted in the highest mean ECD across the seven investigated scenarios for all methods are shown in Table 3. Sensitivity of the results with respect to the choice of tuning parameter values is discussed in the next section.

Table 3: *Optimal prior and tuning parameter values for all methods that resulted in the highest mean ECD across the investigated scenario*

Method	Parameter	Value
BHM	ϕ	0.661
BMA	ψ	-2
CPP	a	2
	b	1.5
CPP-Global	a	1.5
	b	1
	ε	0.5
CPP-Nex	a	2
	b	2
	$\omega^\star$	0.8
EXNEX	ϕ	0.393
	w	0.6
Fujikawa	ε	1.5
	τ	0
JSD-Global	ε	1
	$\varepsilon^\star$	0.5
	τ	0

The results of the comparison regarding the ECDs for each scenario and the mean ECD across all scenarios are shown in Table 4. In terms of the mean ECD, the results are all very similar. All methods lead to mean ECDs between 3.5 and 3.6, differences are only seen in the second decimal place. Note that since there are $K = 4$ baskets the theoretical maximum is 4. The power prior method with CPP-based weights shows the best performance numerically, but only by a very small margin. Numerically the CPP-Nex weights achieved the highest mean ECD and the BMA and MML method showed the lowest values.

Across all scenarios, naturally the best results in terms of ECD are observed in homogeneous scenarios, i.e. when all baskets are either active or inactive, since in these cases information sharing has the highest benefit. Therefore, in the Global Alternative scenario all methods except for MML-Global show more than 3.8 ECD, and in the One in the Middle scenario all methods - excluding the two methods with MML weights - have on average more than 3.7 correct decisions. The scenario in which all methods showed the lowest ECD is the Linear scenario, where all methods have an ECD of around 3. This is because this scenario includes an active basket with a true response probability of only 0.25 and thus only 0.1 above the

Table 4: *ECDs under all scenarios using the optimal tuning parameter values for each method. The best result per column is bold.*

Method	Global Null	Global Alt	One in the Middle	Linear	Good Nugget	Bad Nugget	Half	Mean
BHM	3.928	3.865	3.734	2.999	3.442	3.468	3.365	3.543
BMA	3.904	3.871	3.719	2.964	3.342	3.451	3.319	3.510
CPP	3.916	3.910	3.817	3.066	3.403	3.497	3.321	3.561
CPP-Global	3.922	3.909	3.819	3.056	3.410	3.486	3.323	3.561
CPP-Nex	3.919	3.910	3.816	3.066	3.420	3.494	3.336	**3.566**
EXNEX	3.917	**3.922**	**3.834**	**3.083**	3.343	3.515	3.243	3.551
Fujikawa	3.908	3.882	3.738	3.068	3.340	3.520	3.352	3.544
JSD-Global	3.913	3.919	3.821	3.078	3.392	3.500	3.297	3.560
MML	3.923	3.807	3.624	2.990	3.431	3.516	3.370	3.523
MML-Global	**3.932**	3.640	3.469	2.985	**3.489**	**3.528**	**3.527**	3.510

null response probability. Therefore, low power is expected for this basket, even with two baskets with a higher true response probability.

Looking at the results of the different methods, the BHM design is neither the best nor the worst method in any of the scenarios or in terms of the mean ECD. The BMA design is in the shared last place in terms of mean ECD, and also doesn't show noticeable results in any of the scenarios. The EXNEX design has numerically the best performance in three scenarios, in the Global Alternative, the One in the Middle and the Linear scenario, which indicates that much information is shared with the selected parameters. This leads, however, to below average performance in the Good Nugget scenario and to the worst performance of all methods in the Half scenario.

The three power prior variations with CPP-based weights all have very similar performance in all scenarios. Adding a fixed weight led to a minimal improvement, results with the CPP-Global method, however, are almost indistinguishable from the results of the CPP weights. The CPP based methods are close to the EXNEX design in the Global Alternative, One in the Middle and Linear Scenario, but show little higher ECDs in the Good Nugget and the Half scenario.

Comparing Fujikawa's design and the JSD-Global method, adding the global weights based on the JSD had a bigger influence on the results than adding the global weights based on

h on the results of the CPP method. On average, the JSD-Global method performs a little better than Fujikawa's design, the highest increase in ECD is seen in the One in the Middle scenario. However, Fujikawa's design has a higher ECD than the power prior design with JSD-Global weights in the Bad Nugget and the Half scenario.

The method based on pairwise MML weights is also unremarkable and has average performance in most scenarios, the most noticeable results are seen in the Bad Nugget and the Half scenario, where it is tied with Fujikawa's design for the second place and the sole second place, respectively. The MML-Global method, however, shows the highest ECD in the Good Nugget, Bad Nugget and Half scenario, where the lead in the Half scenario is remarkable, as it beats the second best method by 0.15 ECD, which is the highest lead by any method in any scenario. On the other hand, it takes the last place in the Global Alternative and the One in the Middle scenario, by a similarly large margin.

Rejection rates and FWERs for scenarios with at least one true null hypothesis are shown in Table 5. Note that the FWER in the Global Null scenario is slightly above the nominal level for the BHM design although λ was chosen such that the significance level is protected at 0.05 under the global null hypothesis. Although the simulated data sets are fixed, the MCMC sampling was repeated in the calculation of the rejection rates which led to a slight increase in the estimated FWER.

Table 5: *Rejection rates under all scenarios using the optimal tuning parameter values for each method*

Scenario	Method	Basket 1	Basket 2	Basket 3	Basket 4	FWER
Global Null	BHM	0.020	0.018	0.020	0.018	0.052
	BMA	0.024	0.023	0.024	0.024	0.049
	CPP	0.021	0.021	0.021	0.021	0.048
	CPP-Global	0.019	0.019	0.019	0.019	0.048
	CPP-Nex	0.020	0.020	0.020	0.020	0.049
	EXNEX	0.022	0.020	0.021	0.020	0.049
	Fujikawa	0.023	0.023	0.023	0.023	0.048
	JSD-Global	0.023	0.021	0.023	0.021	0.049
	MML	0.019	0.019	0.019	0.019	0.042
	MML-Global	0.018	0.016	0.017	0.017	0.049
Global Alt	BHM	0.965	0.969	0.966	0.968	.
	BMA	0.967	0.970	0.968	0.967	.

	CPP	0.977	0.977	0.977	0.977	.
	CPP-Global	0.977	0.977	0.977	0.977	.
	CPP-Nex	0.978	0.978	0.978	0.978	.
	EXNEX	0.980	0.982	0.979	0.980	.
	Fujikawa	0.970	0.970	0.970	0.970	.
	JSD-Global	0.980	0.981	0.979	0.980	.
	MML	0.952	0.952	0.952	0.952	.
	MML-Global	0.907	0.912	0.910	0.910	.
One in the Middle	BHM	0.960	0.958	0.826	0.995	.
	BMA	0.955	0.955	0.812	0.996	.
	CPP	0.972	0.972	0.877	0.996	.
	CPP-Global	0.972	0.972	0.878	0.996	.
	CPP-Nex	0.971	0.971	0.877	0.996	.
	EXNEX	0.979	0.978	0.879	0.998	.
	Fujikawa	0.959	0.959	0.824	0.996	.
	JSD-Global	0.975	0.973	0.876	0.997	.
	MML	0.936	0.936	0.760	0.992	.
	MML-Global	0.906	0.904	0.673	0.980	.
Linear	BHM	0.194	0.483	0.781	0.928	0.194
	BMA	0.240	0.492	0.781	0.931	0.240
	CPP	0.247	0.566	0.805	0.942	0.247
	CPP-Global	0.245	0.558	0.805	0.939	0.245
	CPP-Nex	0.248	0.564	0.808	0.942	0.248
	EXNEX	0.284	0.597	0.828	0.938	0.284
	Fujikawa	0.236	0.553	0.807	0.944	0.236
	JSD-Global	0.276	0.584	0.828	0.942	0.276
	MML	0.184	0.486	0.757	0.931	0.184
	MML-Global	0.092	0.391	0.760	0.926	0.092
Good Nugget	BHM	0.060	0.063	0.066	0.628	0.144
	BMA	0.076	0.077	0.080	0.575	0.152
	CPP	0.075	0.075	0.075	0.629	0.154
	CPP-Global	0.072	0.072	0.072	0.627	0.152
	CPP-Nex	0.077	0.077	0.077	0.651	0.161
	EXNEX	0.085	0.086	0.089	0.607	0.174
	Fujikawa	0.087	0.087	0.087	0.602	0.178
	JSD-Global	0.086	0.088	0.088	0.655	0.174
	MML	0.070	0.070	0.070	0.642	0.147
	MML-Global	0.059	0.060	0.061	0.669	0.159
Bad Nugget	BHM	0.272	0.910	0.915	0.915	0.272
	BMA	0.269	0.904	0.907	0.908	0.269
	CPP	0.322	0.940	0.940	0.940	0.322
	CPP-Global	0.322	0.936	0.936	0.936	0.322
	CPP-Nex	0.323	0.939	0.939	0.939	0.323
	EXNEX	0.338	0.947	0.950	0.949	0.338
	Fujikawa	0.288	0.936	0.936	0.936	0.288

	JSD-Global	0.341	0.945	0.949	0.946	0.341
	MML	0.217	0.911	0.911	0.911	0.217
	MML-Global	0.116	0.881	0.880	0.883	0.116
Half	BHM	0.139	0.134	0.821	0.817	0.224
	BMA	0.158	0.157	0.818	0.816	0.222
	CPP	0.179	0.179	0.839	0.839	0.278
	CPP-Global	0.173	0.173	0.835	0.835	0.270
	CPP-Nex	0.178	0.178	0.846	0.846	0.276
	EXNEX	0.220	0.221	0.841	0.837	0.332
	Fujikawa	0.176	0.176	0.852	0.852	0.274
	JSD-Global	0.208	0.208	0.859	0.855	0.311
	MML	0.139	0.139	0.825	0.825	0.225
	MML-Global	0.080	0.079	0.844	0.843	0.144

Results in the Global Alternative scenario are very similar for most methods. All methods expect MML-Global have power values between 0.95 and 0.98 in all baskets. As was already seen in Table 4, this shows that the MML-Global method shares less information than other designs, even in homogeneous scenarios, which results in power values of only around 0.91 in the Global Alternative scenario. The picture is similar in the One in the Middle scenario, though there is more variation in the third basket which has a true response probability of 0.3. The EXNEX design, the CPP-based methods and the JSD-Global method have the highest power in this basket with around 0.87, but other methods only achieve power values that are at least 5 percentage points lower. Again, the two MML-based methods take the last two positions, and MML-Global is far behind with a power of only 0.67.

Looking at the scenarios in which at least one basket is truly inactive, the FWERs are considerably greater than the nominal significance level of 0.05 in most cases. Most methods show the highest type 1 error inflation in the Bad Nugget scenario, where the EXNEX design, the CPP-based methods and the JSD-Global method have FWERs of 0.3 or more. The BHM, BMA and Fujikawa's design still have FWERs of more than 0.25 and MML weights led to a FWER of 0.22. Only with MML-Global weights the type 1 error inflation is more moderate with an FWER of 0.12. However, as expected, the power is also lower and the highest power is achieved by methods with the highest FWER. In the active baskets, the EXNEX design has power of around 0.95, while the power is only 0.88 with MML-Global weights in the power prior design.

In the Good Nugget scenario, there is less variation in the FWER with values between 0.14 and 0.18. Power differences in the active basket are higher though. Here, MML-Global weights show the highest power with 0.67, while the BMA design has the lowest value with 0.58. In this scenario, a large difference between Fujikawa's design and the JSD-Global method is seen, as the latter has a higher power by 5 percentage points while having almost identical FWER.

FWERs are also substantially higher than 5 percent in the Linear scenario with values of up to 0.28 for the EXNEX design and the JSD-Global method. As in the Bad Nugget scenario, the MML-Global method sticks out with an FWER below 10 percent. While the variance of power values in the third basket ($p = 0.35$) and the fourth basket ($p = 0.45$) is moderate, large differences are seen in the second basket, with a true response probability of 0.25. EXNEX and Fujikawa's design as well as the CPP-based methods and the JSD-Global method achieve power values far above 0.5, while with MML-Global weights the power in this basket is below 0.4.

Finally, in the Half scenario, there are again many methods with FWERs above 0.2 and even above 0.3 with the EXNEX design and the JSD-Global method. In terms of power, the JSD-Global method and Fujikawa's design achieve the highest value here, with 0.85, but the CPP-based methods and EXNEX are close behind. Interestingly, the MML-Global method has the second highest power in the two active baskets, although the FWER is once again far lower at only 0.14.

The means of the posterior means are displayed in Figure 7. The green lines correspond to the true response probabilities. As expected, more bias is seen in methods and scenarios where more information is shared with the tuning and prior parameters selected as optimal, and where the true response probabilities are far apart. For example, in the Good Nugget scenario, the mean posterior means for the active basket are between 0.3 and 0.32 for the BHM, BMA and EXNEX design as well as for the CPP-based methods and JSD-Global. Thus the bias is between 8 and 10 percentage points as the true response probability is 0.4. Similarly, in the Bad Nugget scenario, all methods except for the power prior method with MML based weights had mean posterior means above 0.24 in the null basket and thus a bias of at least 9 percentage points. In homogeneous scenarios bias is obviously low.

A mild negative effect of sharing prior information in Fujikawa's design is seen in the Global Null scenario, where its bias amounts to more than 3 percentage points, where for other methods the bias is at most only around 1 percentage point. This is because of the Beta$(1, 1)$ prior, which contains as much information as two patients of which one showed a response. Since all baskets are inactive in this case, much information is shared, thus adding relatively much information through the shared prior parameters. In other scenarios, however, the bias of Fujikawa's design is not relevantly larger than that of the other methods.

3.2.4 Sensitivity Analyses

3.2.4.1 Sensitivity to Choice of Tuning Parameters

Figure 8 illustrates how the mean ECD across all seven scenarios vary, when the tuning and prior parameter values are changed. The choice of the parameter values is obviously important, but for all methods there are a several parameter values that result in high mean ECD, and the results usually do not change drastically when a neighbouring value in the investigated grid is chosen. For the BMA design, large values for ψ which lead to less sharing result in much lower performance in terms of the mean ECD. For the BHM design, very small values for the scale parameter ϕ result in lower mean ECD. In the EXNEX design, interestingly, when holding ϕ fixed, the influence of w is relatively small, unless it is set to 0.1. For the CPP-based methods, there is a large number of parameter value choices for a and b that lead to similar results. Changing $\varepsilon^\star$ with CPP-Global weights has very mild effect. Interestingly, for the CPP-Nex weights, even with w as small as 0.3 good results are achieved, when a and b are changed accordingly. Only with $w = 0.2$ and $w = 0.1$ the mean ECDs decrease relevantly even for adapted a and b. For Fujikawa's design and the power prior design with JSD-Global weights, when the cut-off parameter τ is increased towards 0.5, the performance gets worse for any choice of ε and $\varepsilon^\star$, but relevant decrease in mean ECD is only seen for the highest values in the investigated grid. In Fujikawa's design, small variations in ε are tolerated without too much deterioration in mean ECD. In the power prior design with JSD-Global weights, a similar sensitivity regarding the choice of ε is seen. Higher values of $\varepsilon^\star$ also lead to worse results, but some compensation is possible when smaller values for ε are used.

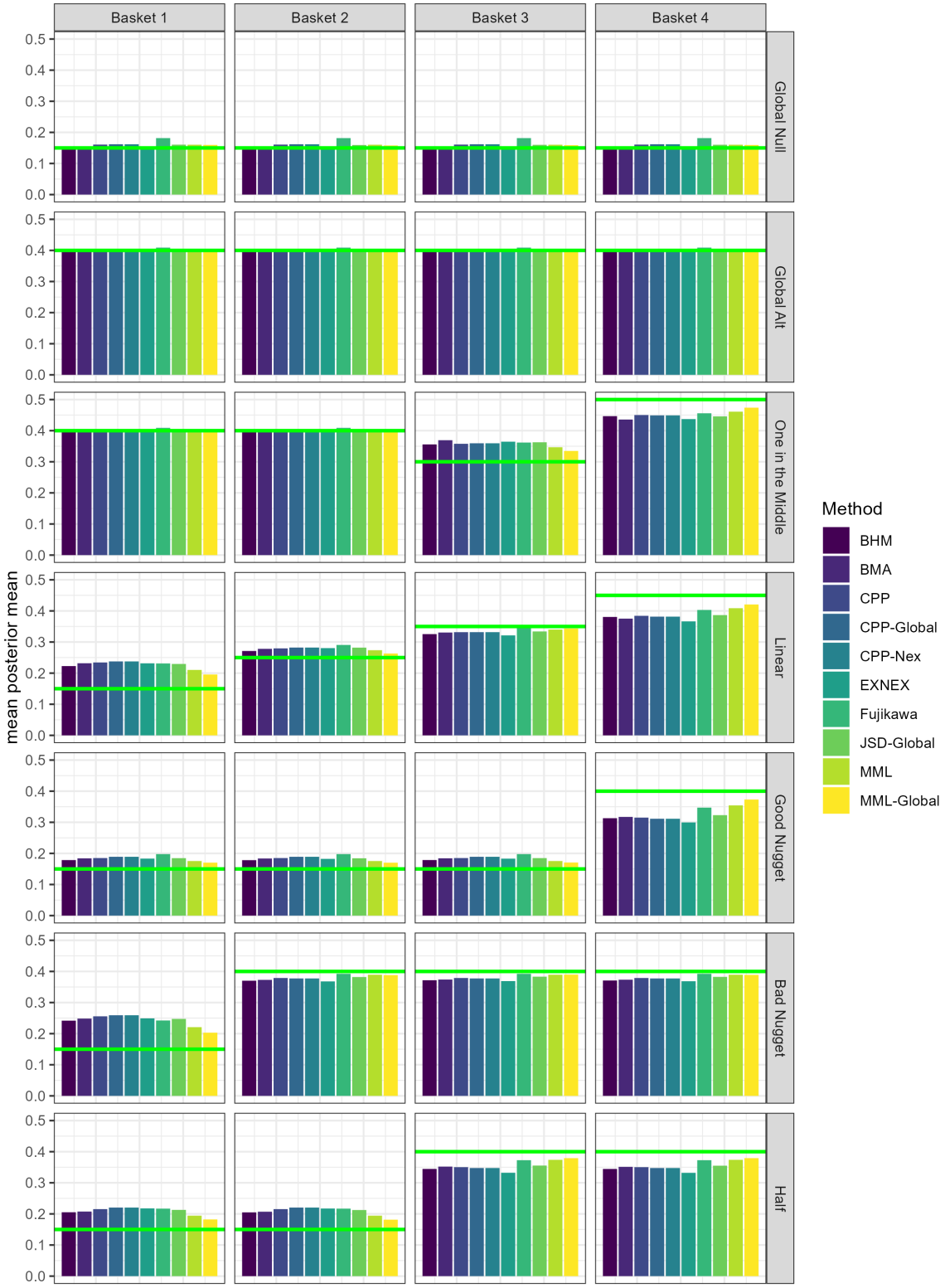

Figure 7: *Mean posterior means and true response probabilities (green lines) in all scenarios with all investigated methods*

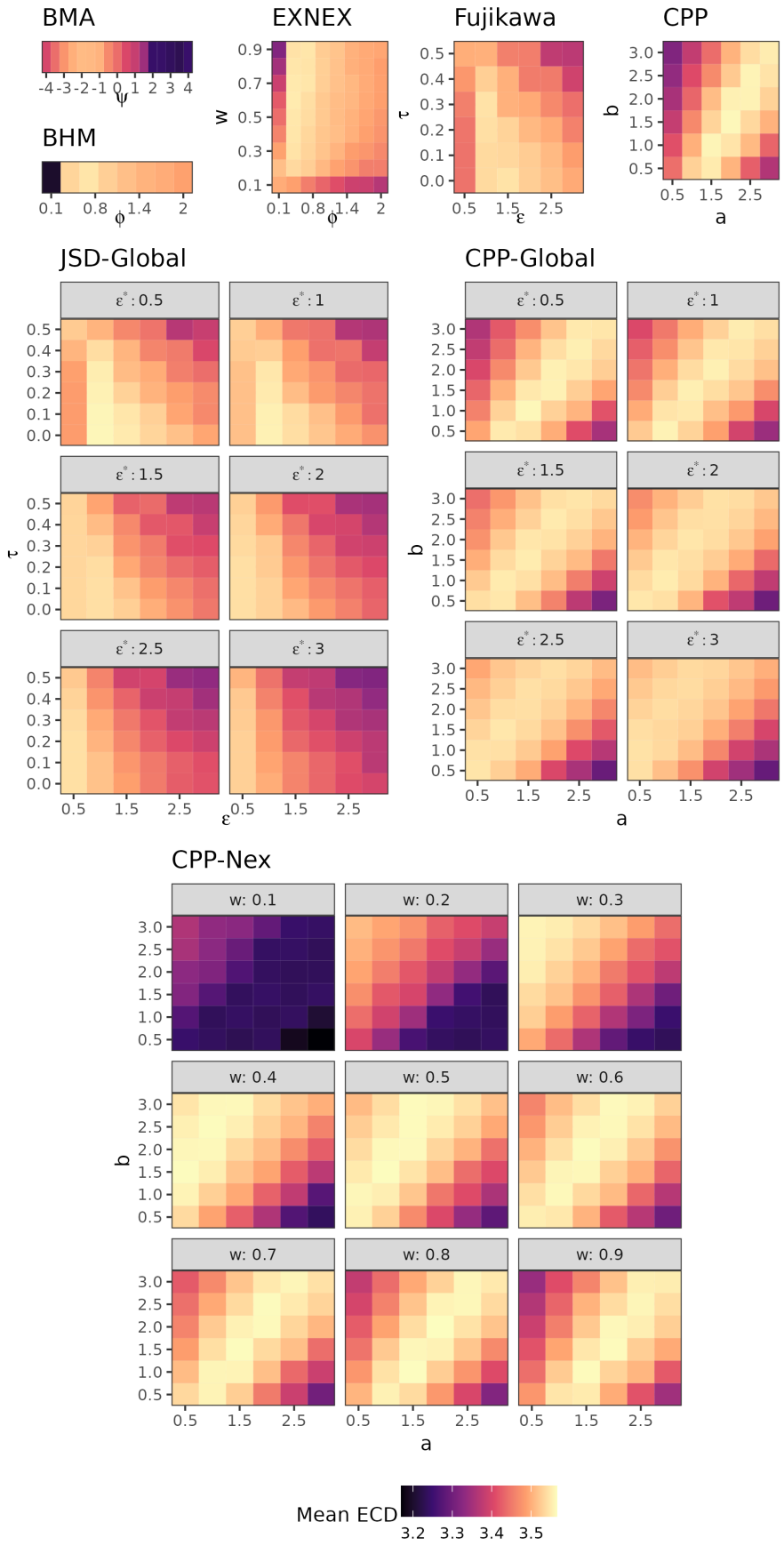

Figure 8: *Influence of prior and tuning parameter values on the results in terms of the mean ECD across the seven scenarios investigated in the comparison study*

3.2.4.2 Sensitivity to the Choice of Scenarios

As a further sensitivity analysis, optimal parameters were chosen based on the mean ECD for only a subset of the seven scenarios. Specifically, the Linear and the One in the Middle scenario were removed from the set of scenarios such that in the remaining scenarios all active baskets have the same response probability of 0.4. This subset of scenarios was used, since in many comparisons studies, such as in Fujikawa et al. (2020), only one alternative response probability is considered in all scenarios.

The results are displayed in Table 6. As a result of removing the Linear scenario, the mean ECDs are higher, but the methods are still very similar. For the limited set of scenarios all methods achieved mean ECDs above 3.6 and again the differences are only seen in the second decimal place. As in the full set of scenarios, the CPP-Nex method has the best numerical mean ECD, but the other two CPP-based methods, the EXNEX and BHM design and also the JSD-Global and MML-Global method are very close behind. Note that since the MML-based methods do not have tuning parameters, the results are the same as in Table 4, but are shown here again for better comparability with the results of the other methods.

Table 6: *ECDs under a subset of scenarios with a single alternative response probability using the optimal tuning parameter values that resulted in the highest mean ECD for each method. The best result per column is bold.*

	Global Null	Global Alt	Good Nugget	Bad Nugget	Half	Mean
BHM	3.938	3.779	3.497	3.501	3.400	3.623
BMA	3.926	3.768	3.465	3.463	3.397	3.604
CPP	3.919	3.779	3.482	3.527	3.433	3.628
CPP-Global	3.930	3.777	**3.503**	3.507	3.410	3.625
CPP-Nex	3.930	**3.813**	3.483	3.520	3.409	**3.631**
EXNEX	**3.937**	3.771	3.498	3.518	3.400	3.625
Fujikawa	3.920	3.780	3.434	**3.540**	3.405	3.616
JSD-Global	3.926	3.783	3.470	3.535	3.406	3.624
MML	3.923	3.807	3.431	3.516	3.370	3.609
MML-Global	3.932	3.640	3.489	3.528	**3.527**	3.623

The tuning parameter values optimal for the reduced set of scenarios in terms of mean ECD are shown in Table 7. Comparing these results with the optimal parameters across all

scenarios in Table 3, changes are seen in all methods but in most cases the difference is small and in many cases the values moved only a few positions in the investigated value-grid. In all cases the changes are seen towards values that share less information, since the results in the two scenarios that were ignored for the sensitivity analysis were better when more information was shared.

Table 7: *Optimal prior and tuning parameter values for the reduced set of scenarios for all methods. Value refers to the tuning parameter values achieving the highest mean ECD across the selected scenarios with a common alternative response probability.*

Method	Parameter	Value
BHM	ϕ	0.929
BMA	ψ	-1
CPP	a	2.5
	b	1.5
CPP-Global	a	2
	b	1
	ε	0.5
CPP-Nex	a	2.5
	b	2.5
	$\omega^\star$	0.6
EXNEX	ϕ	0.929
	w	0.7
Fujikawa	ε	2
	τ	0.1
JSD-Global	ε	2
	$\varepsilon^\star$	0.5
	τ	0

It was further explored how the methods can be tuned when only one specific scenario is of interest. For that sensitivity analysis, the Bad Nugget, Half and Linear scenarios were considered, as these are the most interesting ones for this purpose. For the Global Alternative and One in the Middle scenario all baskets are active and thus pooling all results would be the optimal sharing strategy. The Good Nugget scenario has only one active basket, thus in terms of power nothing can be gained by sharing from the other inactive baskets. Therefore, a basket-wise analysis would be optimal.

Results for all scenarios based on the tuning parameters that are optimal for each of the three selected scenarios are shown in Table 8. In the Linear scenario, only minimal improvements

could be achieved by adapting the tuning parameter values. Based on the parameter values optimal across all scenarios, the EXNEX design achieved the highest ECD with 3.083, with the values chosen specifically for the Linear scenario, Fujikawa's design takes the lead with 3.111 ECD, but the resulting mean ECD is in the second to last place. In all methods, the modified tuning parameter values resulted in some decrease in the mean ECD across the seven scenarios. Only the power prior design with CPP, CPP-Global and JSD-Global weights and the BHM design still have mean ECD above 3.5. A similar picture with relatively little improvements is seen for the results tuned in favour of the Bad Nugget scenario. The EXNEX design has the best performance in this scenario, the JSD-Global method is in the second position and numerically has the highest mean ECD. Most methods are now numerically better than the MML-Global method (not shown here since no tuning parameters are involved, see Table 4). Minimal decrease is seen in the mean ECDs, all methods still have values above 3.5. In the results optimal for the Half scenario, larger improvements are seen. While in the results based on tuning parameters optimal across all scenarios ECD values between 3.3 and 3.4 were seen for most methods, now all methods exceed 3.4 ECD. Note that the MML-Global method still takes the lead in this scenario, being the only method with more than 3.5 ECD. However, as a consequence of the larger improvements in this specific scenario, there was more decrease in the mean ECD for many methods. Only the CPP-Nex method is still above 3.5 mean ECD. The BMA design even deteriorates below 3.3 mean ECD and the CPP and CPP-Global methods are below 3.4 mean ECD.

Table 8: *ECDs under all scenarios using the tuning parameter values that are optimal for the Linear, Bad Nugget and Half scenario.*

Method	Global Null	Global Alt	One in the Middle	Linear	Good Nugget	Bad Nugget	Half	Mean
Optimal for Linear scenario								
BHM	3.907	3.951	3.882	3.059	3.302	3.411	3.210	3.532
BMA	3.835	3.959	3.870	3.009	3.003	3.349	3.051	3.439
CPP	3.908	3.928	3.840	3.088	3.343	3.484	3.293	3.555
CPP-Global	3.890	3.976	3.935	3.071	3.225	3.338	3.097	3.504
CPP-Nex	3.891	3.979	3.942	3.078	3.208	3.322	3.046	3.495
EXNEX	3.884	3.926	3.818	3.101	3.204	3.525	2.966	3.489
Fujikawa	3.878	3.973	3.916	3.111	3.124	3.356	2.996	3.479
JSD-Global	3.89	3.965	3.892	3.093	3.237	3.378	3.079	3.505

Optimal for Bad Nugget scenario								
BHM	3.938	3.779	3.618	2.955	3.497	3.501	3.400	3.527
BMA	3.926	3.768	3.587	2.920	3.465	3.463	3.397	3.504
CPP	3.919	3.779	3.619	2.985	3.482	3.527	3.433	3.535
CPP-Global	3.922	3.766	3.594	2.963	3.486	3.513	3.438	3.526
CPP-Nex	3.921	3.810	3.649	3.017	3.468	3.535	3.406	3.544
EXNEX	3.895	3.891	3.762	3.089	3.281	3.566	3.111	3.514
Fujikawa	3.924	3.794	3.611	3.007	3.412	3.541	3.396	3.526
JSD-Global	3.919	3.846	3.687	3.049	3.435	3.542	3.352	3.547
Optimal for Half scenario								
BHM	3.941	3.706	3.513	2.866	3.540	3.438	3.415	3.488
BMA	3.948	3.001	2.858	2.638	3.598	3.244	3.429	3.245
CPP	3.935	3.388	3.183	2.733	3.556	3.300	3.451	3.364
CPP-Global	3.937	3.431	3.209	2.735	3.563	3.309	3.448	3.376
CPP-Nex	3.924	3.765	3.596	2.962	3.485	3.510	3.445	3.527
EXNEX	3.941	3.687	3.491	2.872	3.546	3.453	3.417	3.487
Fujikawa	3.931	3.600	3.366	2.895	3.484	3.490	3.435	3.457
JSD-Global	3.93	3.604	3.395	2.888	3.51	3.491	3.447	3.467

Table 9 shows the parameters that are optimal for each method for the three selected scenarios. Large changes are especially seen in the Half scenario, where the optimal parameter values share much less information than with the optimal parameter values across all scenarios.

Table 9: *Optimal tuning parameter values for all methods in the Linear, Bad Nugget and Half Scenario*

Method	Parameter	Linear	Bad Nugget	Half
BHM	ϕ	0.393	0.929	2
BMA	ψ	-3.5	-1	4
CPP	a	2	2.5	3
	b	2	1.5	0.5
CPP-Global	a	1.5	3	2.5
	b	2.5	2	0.5
	ε	0.5	0.5	2
CPP-Nex	a	1.5	2.5	3
	b	3	2	2
	$\omega^\star$	0.8	0.8	0.9
EXNEX	ϕ	0.125	0.125	2
	w	0.3	0.2	0.6
Fujikawa	ε	0.5	2	3
	τ	0.4	0	0.2
JSD-Global	ε	0.5	1	3
	$\varepsilon^\star$	0.5	1	0.5
	τ	0.4	0.3	0.2

3.2.5 Summary of the Comparison Study

To summarise the results of the comparison study, on average across all seven scenarios, all methods performed relatively similar. In certain scenarios, in terms of ECD and also rejection rates, some differences were seen, but obviously there is the usual trade-off between between TOER and power, such that methods that had high power in active baskets with the selected prior and tuning parameter values also tended to have high type 1 error inflation. The only method that stood out in that aspect in some scenarios was the power prior design with MML-Global weights, since it had much lower type 1 error inflation than other methods in the Bad Nugget and the Half scenario, but power was on par with the other methods in the Half scenario and only a few percentage points below some other methods in the Bad Nugget scenario. However, the MML-Global methods has non-symmetric sharing weights as explained in Section 3.1.3.2, which is a debatable property. Adding the heterogeneity weights resulted in slightly better results with the JSD-Global method as compared to Fujikawa's design, but adding global weights based on the heterogeneity function h did not improve the power prior design with CPP weights. Overall, the added complexity of the global weights

had no relevant benefit. The sensitivity analyses showed that the ECDs in specific scenarios can be improved when the methods are tuned towards these scenarios, but this leads to a decrease in the ECD in other scenarios. However, the different weights for the power prior design allow a better fine tuning to still have a good performance across all scenarios.

3.3 Nonmonotonic Decisions in Basket Trials

Information sharing between baskets based on observed similarity increases the power, but can also lead to undesirable and unexpected consequences. In this section it will be shown that when the sharing weights depend on the similarity between the subgroups, the posterior probabilities for the response probabilities p_k, $k \in \{1, \ldots, K\}$ may become nonmonotonic in the number of observed responses, which has unexpected and possibly undesired consequences regarding the rejection of null hypotheses in two different ways. First, the null hypothesis of a basket may be rejected even when another basket in the same trial with a higher number of responses is not rejected. Second, one or several null hypotheses may be rejected in a trial, but in another trial with the same number of baskets and the same sample sizes, a uniformly higher number of responses can lead to no rejected null hypotheses.

Kopp-Schneider et al. (2020) describe a closely related issue in the context of borrowing information from a single external study. They consider an example of a single-arm trial with a sample size of $n = 40$ and a binary outcome where $r = 12$ responses were observed. The null response rate is $p_0 = 0.2$. It is further assumed that $n_0 = 100$ external data points are available and in this sample $r_0 = 30$ responses were seen. Hence, the response rates are identical in the current and the external study. The authors consider, among others, an "extreme borrowing" approach where 100% of the external information is shared (i.e. the data are pooled) if and only if the response rates of both data sets are identical. Using a probability threshold of $\lambda = 0.9976$ and rejecting the null hypothesis if $\mathbb{P}(p > p_0|r, r_0) > \lambda$ in this case leads to a one-sided TOER of 0.047, i.e. the one-sided significance level is protected at 0.05. However, the extreme borrowing approach results in a posterior probability $\mathbb{P}(p > p_0|r, r_0)$ that is not monotonic in r and as a consequence even in a non-connected rejection region: The null hypothesis is rejected if $r \in \{12, 16, 17, 18, \ldots, 39, 40\}$.

Clearly, as the authors note, a non-connected rejection region would not be used in a real clinical trial. In the context of borrowing information from a single external data source in a single-arm trial, this issue is relatively easy to identify and generally does not occur when more moderate borrowing approaches are used. In basket trials, however, information is borrowed not only from one source, but from several other baskets. Furthermore, K hypotheses are tested instead of one. This can lead to nonmonotonic decisions similar to the example given above as is shown in the next sections. More specifically, there are two different kinds of nonmonotonicity. The first (termed within-trial nonmonotonicity) is seen when test decisions of the baskets in a single trial are compared. The second (termed between-trial nonmonotonicity) occurs between the outcomes of different trials.

In the following, both types of nonmonotonicity are at first demonstrated using specific examples. Then, nonmonotonicity is investigated systematically in Fujikawa's design and the power prior design with CPP weights.

3.3.1 Within-Trial Nonmonotonicity

Consider as an example a single-stage basket trial with $K = 4$ baskets, a null response rate of $p_0 = 0.15$ and $n = 20$ patients per basket. The power prior design with CPP weights is used to analyse the results. The tuning parameters are set to $a = 1.5$ and $b = 0.5$, the prior for each basket is $\text{Beta}(1, 1)$. The probability threshold is set to $\lambda = 0.99$. Let the observed outcome vector be $\boldsymbol{r} = (5, 5, 5, 6)$. Without sharing, the posterior distributions and probabilities are:

$$\pi(p_i|r_i = 5) = \text{Beta}(6, 16), \quad \mathbb{P}(p_i > 0.15|r_i = 5) = 0.92, \text{for } i \in \{1, 2, 3\},$$
$$\pi(p_4|r_4 = 6) = \text{Beta}(7, 15), \quad \mathbb{P}(p_4 > 0.15|r_4 = 6) = 0.97.$$

Thus, if the baskets are analysed individually, no null hypothesis can be rejected. Clearly, the posterior probabilities for $p_k > p_0$ are strictly monotonically increasing in r_k. With the above specified parameters for information sharing, the baskets with 5 responses share 100% of the information with each other. Between the baskets with 5 responses and the basket with 6 responses the sharing weight is 0.41. This leads to the posterior distributions and

probabilities as follows:

$$\pi(p_i|\boldsymbol{r}, \boldsymbol{\omega_i}) = \text{Beta}(18.4, 51.7), \quad \mathbb{P}(p_i > 0.15|\boldsymbol{r}, \boldsymbol{\omega_i}) = 0.992, \text{for } i \in \{1, 2, 3\},$$
$$\pi(p_4|\boldsymbol{r}, \boldsymbol{\omega_4}) = \text{Beta}(13.1, 33.3), \quad \mathbb{P}(p_4 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_4}) = 0.988.$$

Hence, for the three baskets with 5 responses the null hypothesis of $p_0 = 0.15$ can be rejected. For the fourth basket with 6 responses it cannot be rejected, although there was one additional responder. This happens because the first three baskets share all of their information due to the identical response rates, while the first basket only shares 41% of the information with baskets 1 to 3. Note that the expected value of the posterior distribution of basket 4 is 0.28, which is greater than that of the first three baskets with 0.26. However, the variance is also greater - $4.27 \cdot 10^{-3}$ vs. $2.72 \cdot 10^{-3}$ - which leads to this result.

When there is prior evidence that the treatment under investigation leads to similar response rates in a subset of the baskets, then it may be reasonable that efficacy is declared for these baskets even when the probability threshold is not reached for another basket with a higher response rate. However, when such information is not available, then this seems like a very undesirable result.

3.3.2 Between-Trial Nonmonotonicity

Consider now again a single-stage basket trial with the same K, p_0 and n as above. Now, the power prior design with MML weights is used for the analysis, again with Beta$(1, 1)$ priors. The probability threshold is now set to $\lambda = 0.97$. At first, let the observed outcome vector be $\boldsymbol{r} = (0, 1, 5, 6)$. The posterior distributions and probabilities with shared information are as follows:

$$\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) = \text{Beta}(2.6, 40.9), \quad \mathbb{P}(p_1 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.021,$$
$$\pi(p_2|\boldsymbol{r}, \boldsymbol{\omega_2}) = \text{Beta}(3.3, 42.9), \quad \mathbb{P}(p_2 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_2}) = 0.039,$$
$$\pi(p_3|\boldsymbol{r}, \boldsymbol{\omega_3}) = \text{Beta}(12.2, 34.5), \quad \mathbb{P}(p_3 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_3}) = 0.971,$$
$$\pi(p_4|\boldsymbol{r}, \boldsymbol{\omega_4}) = \text{Beta}(12.1, 32.6), \quad \mathbb{P}(p_4 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_4}) = 0.978.$$

Thus, with information sharing two of the four null hypotheses for the two baskets with 5 and 6 responses can be rejected. In a basket-wise analysis only the posterior probability of the fourth basket would reach the threshold ($\mathbb{P}(p_4 > 0.15|r_4 = 6) = 0.971$), but basket 3 borrows all information from basket 4 and very little information from the other two baskets, thus also resulting in a posterior probability for which the null hypothesis is rejected.

Now assume that in the second basket 2 instead of 1 responses were observed, hence the response vector is $\boldsymbol{r} = (0, 2, 5, 6)$. The following posterior distributions and probabilities are then observed:

$$\begin{aligned}
\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) &= \text{Beta}(3.0, 35.5), \quad \mathbb{P}(p_1 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.068, \\
\pi(p_2|\boldsymbol{r}, \boldsymbol{\omega_2}) &= \text{Beta}(6.7, 43.5), \quad \mathbb{P}(p_2 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_2}) = 0.332, \\
\pi(p_3|\boldsymbol{r}, \boldsymbol{\omega_3}) &= \text{Beta}(13.0, 40.3), \quad \mathbb{P}(p_3 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_3}) = 0.958, \\
\pi(p_4|\boldsymbol{r}, \boldsymbol{\omega_4}) &= \text{Beta}(12.4, 34.5), \quad \mathbb{P}(p_4 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_4}) = 0.975.
\end{aligned}$$

The additional response in basket 2 has the consequence that more information is shared between this and the third and fourth basket, thus pulling the posterior mean of these two baskets closer to the null response rate. The posterior mean of the third basket was 0.261 before and is now 0.244, that of the fourth basket was 0.271 and with the additional response in the second basket is now 0.264. While in the basket with 6 responses the null hypothesis can still be rejected, in basket 3 the posterior probability now decreased to 0.958 and thus below the posterior threshold of $\lambda = 0.97$. Now, going one step further, assume that 2 more responses were observed, one in basket 1 and one in basket 2, i.e. $\boldsymbol{r} = (1, 3, 5, 6)$. The results are then:

$$\begin{aligned}
\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) &= \text{Beta}(5.7, 37.0), \quad \mathbb{P}(p_1 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.338, \\
\pi(p_2|\boldsymbol{r}, \boldsymbol{\omega_2}) &= \text{Beta}(13.5, 56.6), \quad \mathbb{P}(p_2 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_2}) = 0.817, \\
\pi(p_3|\boldsymbol{r}, \boldsymbol{\omega_3}) &= \text{Beta}(15.2, 50.1), \quad \mathbb{P}(p_3 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_3}) = 0.954, \\
\pi(p_4|\boldsymbol{r}, \boldsymbol{\omega_4}) &= \text{Beta}(14.0, 42.3), \quad \mathbb{P}(p_4 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_4}) = 0.968.
\end{aligned}$$

Now, as a consequence of the increased number of responses in the first two baskets, the heterogeneity of the results is decreased and more information is shared. This further reduces

the posterior means of basket 3 and basket 4 which are now 0.232 and 0.248, respectively. This also has the effect that now even for the fourth basket the null hypothesis can not be rejected with a posterior probability of 0.968. Hence, although more responses were observed in the first two baskets and the identical number of responses was observed in the third and fourth basket, looking only at the rejected null hypotheses, the results changed entirely from rejecting two null hypotheses to rejecting none. This is quite counterintuitive, as the same or more evidence for an effective treatment was observed in all four baskets. Even if the number of responses in the first basket is further increased to 3, the number of rejected null hypotheses is still zero.

Even more extreme changes in the number of rejected null hypotheses may result from a minimal change in the results. With the same design specifications as above, assume $\boldsymbol{r} = (1, 5, 5, 5)$ to obtain the following results:

$$\begin{aligned}
&\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) = \text{Beta}(4.5, 27.4), \quad \mathbb{P}(p_1 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.390, \\
&\pi(p_i|\boldsymbol{r}, \boldsymbol{\omega_i}) = \text{Beta}(16.2, 49.1), \quad \mathbb{P}(p_i > 0.15|\boldsymbol{r}, \boldsymbol{\omega_i}) = 0.977, \text{ for } i \in \{2, 3, 4\}.
\end{aligned}$$

Thus, three null hypotheses can be rejected for the three baskets in which 5 responses were observed. But only adding one response to the the first basket, i.e. $\boldsymbol{r} = (2, 5, 5, 5)$, leads to an entirely different conclusion in terms of rejected null hypotheses:

$$\begin{aligned}
&\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) = \text{Beta}(10.5, 41.4), \quad \mathbb{P}(p_1 > 0.15|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.823, \\
&\pi(p_i|\boldsymbol{r}, \boldsymbol{\omega_i}) = \text{Beta}(17.0, 54.9), \quad \mathbb{P}(p_i > 0.15|\boldsymbol{r}, \boldsymbol{\omega_i}) = 0.969, \text{ for } i \in \{2, 3, 4\}.
\end{aligned}$$

Now, no null hypothesis can be rejected.

In other cases, however, nonmonotonicity can look more intuitive. Let now $p_0 = 0.3$ and $\lambda = 0.99$. The power prior design with CPP weights and $a = 2.5$ and $b = 3$ is used to analyse the outcome vector $\boldsymbol{r} = (0, 0, 10, 10)$. The results are then:

$$\begin{aligned}
&\pi(p_i|\boldsymbol{r}, \boldsymbol{\omega_i}) = \text{Beta}(2.3, 42.3), \quad \mathbb{P}(p_i > 0.3|\boldsymbol{r}, \boldsymbol{\omega_i}) < 0.001 \text{ for } i \in \{1, 2\}, \\
&\pi(p_j|\boldsymbol{r}, \boldsymbol{\omega_j}) = \text{Beta}(21.0, 23.6), \quad \mathbb{P}(p_j > 0.3|\boldsymbol{r}, \boldsymbol{\omega_j}) = 0.991, \text{ for } j \in \{3, 4\}.
\end{aligned}$$

Since the response rates are identical in basket 1 and basket 2 as well as in basket 3 and basket 4, the data in these baskets are fully shared. Between the baskets with different response rates, very little information is shared due to the highly different response rates. Because of the information sharing, the null hypotheses are rejected in the baskets with 10 responses (if no information was shared, 11 responses would be necessary for a posterior probability above 0.99). Now compare this to the results obtained when the outcome vector is $\boldsymbol{r} = (5, 7, 10, 10)$:

$$\begin{aligned}
&\pi(p_1|\boldsymbol{r}, \boldsymbol{\omega_1}) = \text{Beta}(19.4, 34.8), \quad \mathbb{P}(p_1 > 0.3|\boldsymbol{r}, \boldsymbol{\omega_1}) = 0.813,\\
&\pi(p_2|\boldsymbol{r}, \boldsymbol{\omega_2}) = \text{Beta}(26.9, 41.9), \quad \mathbb{P}(p_2 > 0.3|\boldsymbol{r}, \boldsymbol{\omega_2}) = 0.943,\\
&\pi(p_j|\boldsymbol{r}, \boldsymbol{\omega_j}) = \text{Beta}(27.8, 35.7), \quad \mathbb{P}(p_i > 0.3|\boldsymbol{r}, \boldsymbol{\omega_j}) = 0.989, \text{ for } j \in \{3, 4\}.
\end{aligned}$$

Due to the better outcome in baskets 1 and 2, more information is shared overall and as a consequence the null hypotheses for basket 3 and 4 cannot be rejected anymore, although the number of responses remains the same in these two baskets and 12 more responses were observed in total. In this case, however, the nonmonotonicity seems more reasonable. For the first outcome vector with 0 responses in the first two baskets, it seems natural to conclude that there are two different basket clusters. In the first cluster with basket 1 and basket 2, the treatment is futile and in the second cluster comprising basket 3 and basket 4, the treatment looks effective. When in the second case the number of responses is increased to 5 and 7 in the first two baskets, the results indicate more similar response rates in all baskets, with some variation in the observed responses. Hence, sharing a high percentage of information across all baskets seems appropriate. Of course this results in lower posterior means in the baskets with a higher number of responses and in this case also in a posterior probability that drops below the probability threshold.

3.3.3 Monotonicity Conditions

In the last sections, two types of nonmonotonicity that can appear as a consequence of sharing information between baskets based on the observed similarity of their outcomes were shown. The first type of nonmonotonicity appears within the outcome of a single trial, when the null hypothesis is rejected for a basket, but it is not rejected for another basket in the trial for which more responses were observed. The second type is seen when two outcome vectors

are compared, where a uniformly higher number of responses across all baskets can reduce the number of rejected null hypotheses to 0. When these two types of nonmonotonicity are considered unacceptable, the following monotonicity conditions are suggested for a setting with equal sample sizes in all baskets:

Within-Trial Monotonicity Condition: If a null hypothesis $H_{0,k}$ is rejected for a basket k, $k \in \{1, \ldots, K\}$, in which r_k responses were observed, then for any basket j, $j \in \{1, \ldots, K\}$ for which the number of responses is larger than or equal to the number of responses in basket k, i.e. $r_j \geqslant r_k$, the null hypothesis $H_{0,j}$ must also be rejected.

Between-Trials Monotonicity Condition: If at least one null hypothesis $H_{0,k}$, $k \in \{1, \ldots, K\}$ can be rejected if the vector of responses is $\boldsymbol{r} = (r_1, \ldots, r_K)$, then at least one null hypothesis $H_{0,j}$, $j \in \{1, \ldots, K\}$ must also be rejected for all response vectors $\boldsymbol{r'} = (r'_1, \ldots, r'_K)$ for which $r'_{(i)} \geqslant r_{(i)}$ holds for all $i \in \{1, \ldots, K\}$, where $r_{(i)}$ refers to the i-th element of the vector $\boldsymbol{r}$ with sorted elements.

The second condition is formulated rather weakly and a stronger version could be used instead. It may seem more natural to require that the same number of null hypotheses is rejected for the outcome vectors $\boldsymbol{r}$ and $\boldsymbol{r'}$ when $r'_{(i)} \geqslant r_{(i)}$ holds for all $i \in \{1, \ldots K\}$. However, this stronger condition would only hold when information is shared even between baskets for which the response rates are far apart. To illustrate this, consider a basket trial with only $K = 2$ baskets with $p_0 = 0.15$, $n = 20$ per basket and Beta$(1, 1)$ priors in both baskets and $\lambda = 0.99$. For the analysis, a power prior based design with pairwise weights is used, hence information is fully shared if the number of responses is identical in the two baskets. If a separate analysis is conducted in each basket and no information is shared, 7 response are necessary to reject the null hypothesis. If $\boldsymbol{r} = (6, 6)$ is observed and information sharing is used, then both null hypotheses can be rejected. Now assume that the outcome vector $\boldsymbol{r'} = (6, 20)$ is observed. Clearly, for any choice of sharing weights the null hypothesis for the second basket can be rejected, so this does not violate the between-trials monotonicity condition as formulated above. If it would be required, however, that the first null hypothesis is also rejected, then necessarily some information has to be shared between these two baskets despite their very different response rates, which seems too strong as a requirement. Some of

the investigated methods still share some information in this case - e.g. the CPP method with the selected optimal tuning parameter values results in weights of 0.063 which is still enough to reject the null hypothesis for the first basket. With MML weights, for example, the weight in this case is 0, thus $H_{0,1}$ would not be rejected. The weak condition proposed above seems more reasonable and sufficiently strict, especially in the context of proof of concept studies in which the main purpose is to examine whether the investigated treatment is effective in any of the baskets. The stronger version could of course still be used.

3.3.4 Avoiding Nonmonotonicity by Pruning Baskets

The examples of the previous section demonstrated that nonmonotonicity appears when information sharing between baskets with a different number of responses pulls the posterior mean of the baskets with a better outcome closer towards p_0. A simple approach to prevent nonmonotonicity in many cases is thus to ignore baskets with few responses in the information sharing, which is called pruning in the following. Let $c = c_k$ be the critical value as defined in Equation (2.9), which is identical for all baskets since equal sample sizes and prior distributions are assumed. Trivially, if baskets for which $r_k < c$ are pruned, the monotonicity conditions hold. The null hypotheses are then rejected for baskets with $r_k \geqslant c$, i.e. for all baskets that are not pruned. But clearly there is no power gain from information sharing in this case. Baskets could instead be pruned when r_k is smaller than the "pooled critical value", c_{pool}, which is defined as:

$$c_{\text{pool}} = \min\{r^\star \in \mathbb{N} : \mathbb{P}(p_k > p_0 | (r^\star, \ldots, r^\star), \boldsymbol{\omega_k}) \geqslant \lambda\}.$$

c_{pool} is therefore the smallest integer $r^\star$ for which the null hypotheses of all baskets can be rejected if $r_k = r^\star$ for all $k \in \{1, \ldots, K\}$. Note that this definition only makes sense when sample sizes and prior distributions are equal in all baskets.

Note that c_{pool} decreases with increasing K. For $n = 20$, $p_0 = 0.15$, $\lambda = 0.99$ and with Beta$(1, 1)$ priors, $c_{\text{pool}} = 6$ for the power prior design with $K = 3$ and $c_{\text{pool}} = 5$ for $K = 4$. The next decrease occurs at $K = 15$, where $c_{\text{pool}} = 4$. c_{pool} cannot decrease further, as with 20 observations and 3 responses the observed response rate is 0.15 and thus identical to p_0. Note further that for Fujikawa's design c_{pool} may differ from that of the power prior design,

as in Fujikawa's design the prior information is also shared. Using the same parameters as above, for example, $c_{\text{pool}} = 5$ with 3 baskets in Fujikawa's design.

Pruning baskets when $r_k < c_{\text{pool}}$ resolves between-trial nonmonotonicity in all examples shown in Section 3.3.2. In the first example, where the vector of responses is increased from $(0, 1, 5, 6)$ to $(0, 2, 5, 6)$ and then $(1, 3, 5, 6)$, $c_{\text{pool}} = 5$ and thus for all three vectors information is only shared between the third and fourth basket. With the MML method the information is still fully shared in this case, although the response rates are not identical, which leads to posterior probabilities of 0.985 that are thus greater than λ for both baskets. In the second example where the two response vectors are $(1, 5, 5, 5)$ and $(2, 5, 5, 5)$, the first basket is pruned in both cases and thus there is also no nonmonotonicity. In the third example $p_0 = 0.3$ was assumed and thus the pooled critical value is different, namely $c_{\text{pool}} = 8$. Therefore, in the two vectors $(0, 0, 10, 10)$ and $(5, 7, 10, 10)$ the first two baskets are pruned in both cases. In these three examples, the null hypotheses in the baskets with a number of responses greater than or equal to c_{pool} can still be rejected. Pruning can, however, also lead to cases where no null hypothesis can be rejected as a consequence of pruning. The effect of pruning on power is investigated in Section 3.3.6.

In the within-trial nonmonotonicity example in Section 3.3.1, pruning baskets with $r_k < c_{\text{pool}}$ does not resolve the issue. In this case p_0 was set to 0.15 and thus c_{pool} is again 5. Thus nothing is pruned in the vector $(5, 5, 5, 6)$. The nonmonotonicity would therefore only be resolved if all baskets with less than 6 responses were pruned.

As was shown in this section, pruning baskets with a number of responses smaller than c_{pool} can resolve nonmonotonicity in many cases but not in all. When pruning is decided to be an appropriate strategy, different values may be tried but c_{pool} can be a starting point. Using the smallest possible cut-off for pruning such that nonmonotonicity is resolved seems desirable to avoid losing too much power. Pruning may also be seen as beneficial to prevent null hypotheses in baskets with a very low response rate to be rejected, e.g. when the observed response rate in a basket is below p_0.

In the planning phase of a basket trial, researchers may at first investigate whether any nonmonotonic events can occur with the chosen setup. If this is the case, then it can be

decided whether the nonmonotonicity is acceptable or not. If not, then the pruning strategy can be applied and the influence of different pruning cut-offs may be tested. The pruning cut-off should be prespecified with the rest of the analysis strategy.

3.3.5 Investigation of the Monotonicity Conditions

In this section, it is investigated whether the monotonicity conditions defined in the previous section hold in Fujikawa's design and in the power prior design with CPP weights with different tuning parameter values and with and without pruning baskets with $r_k < c_{\text{pool}}$. $K \in \{4, 5, 6\}$ baskets were considered. As before, the sample size was 20 in each basket and Beta$(1, 1)$ prior distributions were used for both designs. The probability threshold was set to $\lambda = 0.99$. The same tuning parameter values as for the comparison study in Section 3.2 were tested: For the tuning parameters a and b of the CPP method and for ε in Fujikawa's design values between 0.5 and 3 in steps of 0.5 were used and for τ in Fujikawa's design values between 0 and 0.5 in steps of 0.1 were examined.

The results are shown in Table 10. Interestingly, the results without pruning are quite different for the two methods. In Fujikawa's design, the within-trial condition holds for most investigated parameters. With 4 baskets the condition was never violated, for 5 and 6 baskets it was only violated for $\varepsilon = 2.5$ with all choices of τ. The between-trial condition was violated for most parameter choices in 4 baskets and for all examined parameter values for 5 and 6 baskets. With the CPP method, there are many parameter value combinations that do not lead to a violation of the between-trial monotonicity condition. The number of values that does violate the condition increases with K, but even with $K = 6$ the condition still holds for the majority of the considered values of a and b. However, there are much more cases where the within-trial condition is violated. The number of parameter values for which this is the case also increases in K.

Pruning resolves nonmonotonicity in all tested cases in Fujikawa's design. With the CPP method, there are some parameter combinations left which violate the monotonicity conditions. Note, however, that with the parameter values for which is the case - small values of b and larger values of a - the sharing weights decline rapidly when the response rates

differ, see Figure 4. These are cases that are close to the extreme borrowing discussed in Kopp-Schneider et al. (2020).

Table 10: *Results of the investigation of the monotonicity condition in Fujikawa's design and in the power prior design with CPP weights, with and without pruning. A cross means that the monotonicity condition is violated, a tick means that the monotonicity condition holds.*

Fujikawa - Without Pruning													
Within-trial condition (K = 4)							Between-trial condition (K = 4)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✗	✗	✗	✗	✗	✗
1	✓	✓	✓	✓	✓	✓	1	✗	✗	✗	✗	✗	✗
1.5	✓	✓	✓	✓	✓	✓	1.5	✗	✗	✗	✗	✗	✗
2	✓	✓	✓	✓	✓	✓	2	✗	✗	✗	✗	✗	✓
2.5	✓	✓	✓	✓	✓	✓	2.5	✗	✗	✗	✗	✓	✓
3	✓	✓	✓	✓	✓	✓	3	✗	✗	✗	✓	✓	✓
Within-trial condition (K = 5)							Between-trial condition (K = 5)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✗	✗	✗	✗	✗	✗
1	✓	✓	✓	✓	✓	✓	1	✗	✗	✗	✗	✗	✗
1.5	✓	✓	✓	✓	✓	✓	1.5	✗	✗	✗	✗	✗	✗
2	✓	✓	✓	✓	✓	✓	2	✗	✗	✗	✗	✗	✗
2.5	✗	✗	✗	✗	✗	✗	2.5	✗	✗	✗	✗	✗	✗
3	✓	✓	✓	✓	✓	✓	3	✗	✗	✗	✗	✗	✗
Within-trial condition (K = 6)							Between-trial condition (K = 6)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✗	✗	✗	✗	✗	✗
1	✓	✓	✓	✓	✓	✓	1	✗	✗	✗	✗	✗	✗
1.5	✓	✓	✓	✓	✓	✓	1.5	✗	✗	✗	✗	✗	✗
2	✓	✓	✓	✓	✓	✓	2	✗	✗	✗	✗	✗	✗
2.5	✗	✗	✗	✗	✗	✗	2.5	✗	✗	✗	✗	✗	✗
3	✓	✓	✓	✓	✓	✓	3	✗	✗	✗	✗	✗	✗
CPP - Without Pruning													
Within-trial condition (K = 4)							Between-trial condition (K = 4)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✗	✓	✓	✓	✓	✓
1.5	✗	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	2	✗	✓	✓	✓	✓	✗
2.5	✓	✗	✓	✓	✓	✓	2.5	✗	✓	✓	✓	✓	✗
3	✓	✗	✓	✓	✓	✓	3	✗	✗	✓	✗	✓	✗

Within-trial condition (K = 5)							Between-trial condition (K = 5)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✗	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✗	✗	✓	✓	✓	✓	1	✗	✓	✓	✓	✓	✓
1.5	✗	✓	✓	✓	✓	✓	1.5	✗	✓	✓	✓	✓	✓
2	✗	✗	✓	✓	✓	✓	2	✗	✓	✓	✓	✓	✗
2.5	✗	✗	✗	✓	✓	✓	2.5	✗	✗	✓	✓	✗	✗
3	✗	✗	✗	✓	✓	✓	3	✗	✗	✓	✓	✗	✗
Within-trial condition (K = 6)							Between-trial condition (K = 6)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✗	✗	✓	✓	✓	✓	0.5	✗	✓	✓	✓	✓	✓
1	✗	✗	✓	✓	✓	✓	1	✗	✓	✓	✓	✓	✓
1.5	✗	✗	✓	✓	✓	✓	1.5	✗	✓	✓	✓	✓	✓
2	✗	✗	✗	✓	✓	✓	2	✗	✗	✓	✓	✓	✗
2.5	✗	✗	✗	✓	✓	✓	2.5	✗	✗	✓	✓	✗	✗
3	✗	✗	✗	✓	✓	✓	3	✗	✗	✓	✗	✗	✗

Fujikawa - With Pruning

Within-trial condition (K = 4)							Between-trial condition (K = 4)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✓	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	✓
2.5	✓	✓	✓	✓	✓	✓	2.5	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	✓
Within-trial condition (K = 5)							Between-trial condition (K = 5)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✓	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	✓
2.5	✓	✓	✓	✓	✓	✓	2.5	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	✓
Within-trial condition (K = 6)							Between-trial condition (K = 6)						
ε / τ	0	0.1	0.2	0.3	0.4	0.5	ε / τ	0	0.1	0.2	0.3	0.4	0.5
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✓	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	✓
2.5	✓	✓	✓	✓	✓	✓	2.5	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	✓

CPP - With Pruning

Within-trial condition (K = 4)							Between-trial condition (K = 4)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✗	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	2	✗	✓	✓	✓	✓	✓
2.5	✓	✗	✓	✓	✓	✓	2.5	✗	✓	✓	✓	✓	✓
3	✓	✗	✓	✓	✓	✓	3	✗	✓	✓	✓	✓	✓
Within-trial condition (K = 5)							Between-trial condition (K = 5)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✗	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✗	✓	✓	✓	✓	✓	2	✗	✓	✓	✓	✓	✓
2.5	✗	✗	✓	✓	✓	✓	2.5	✗	✓	✓	✓	✓	✓
3	✗	✗	✓	✓	✓	✓	3	✗	✓	✓	✓	✓	✓
Within-trial condition (K = 6)							Between-trial condition (K = 6)						
a / b	0.5	1	1.5	2	2.5	3	a / b	0.5	1	1.5	2	2.5	3
0.5	✓	✓	✓	✓	✓	✓	0.5	✓	✓	✓	✓	✓	✓
1	✓	✓	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	✓
1.5	✗	✓	✓	✓	✓	✓	1.5	✓	✓	✓	✓	✓	✓
2	✗	✓	✓	✓	✓	✓	2	✗	✓	✓	✓	✓	✓
2.5	✗	✗	✓	✓	✓	✓	2.5	✗	✓	✓	✓	✓	✓
3	✗	✗	✓	✓	✓	✓	3	✗	✓	✓	✓	✓	✓

3.3.6 Influence of Pruning on the Operating Characteristics

In the previous section it was shown that pruning baskets with $r_k < c_{\text{pool}}$ ensures that the monotonicity conditions hold for most choices of tuning parameter values. The effect of pruning on the operating characteristics is, however, not obvious. To investigate this, for Fujikawa's design and the power prior method with CPP weights the comparison study as presented in Section 3.2 was performed again, but with pruning. Tuning parameter values were again selected based on the highest mean ECD while controlling the FWER under the global null hypothesis at 5%. Results in terms of the ECDs in all scenarios are shown in Table 11. Results without pruning are included again for better comparison, but are the same as seen in Table 4.

Table 11: *ECDs of Fujikawa's design and the power prior design with CPP weights with and without pruning under all scenarios using the optimal tuning parameter values*

Method	Global Null	Global Alt	One in the Middle	Linear	Good Nugget	Bad Nugget	Half	Mean
CPP	3.916	3.910	3.817	3.066	3.403	3.497	3.321	3.561
CPP (pruning)	3.930	3.495	3.310	2.973	3.457	3.543	3.553	3.466
Fujikawa	3.908	3.882	3.738	3.068	3.340	3.520	3.352	3.544
Fujikawa (pruning)	3.919	3.792	3.648	3.146	3.383	3.644	3.393	3.561

The optimal parameter values for Fujikawa's design are $\varepsilon = 1$ and $\tau = 0$. For the power prior design with CPP methods, 17 of the investigated 36 parameter value combinations resulted in exactly the same mean ECD. The results are displayed in Table 11. The difference regarding the effect of pruning on the results of the two methods is quite large. While the mean ECD of the CPP method decreased from 3.56 to 3.47 with Fujikawa's design it increased from 3.54 to 3.56. Interestingly, this difference seems to be a result of including the prior parameters in the information sharing in Fujikawa's design. Indeed, when Fujikawa's design is modified so that the prior parameters are not shared - so a power prior design with JSD weights - the results of the optimisation are almost identical to those of the CPP method with pruning.

Table 12: *Rejection rates of the power prior design with CPP weights and Fujikawa's design with and without pruning, under all scenarios using the optimal tuning parameter values for each method*

Scenario	Method	Basket 1	Basket 2	Basket 3	Basket 4	FWER
Global Null	CPP	0.021	0.021	0.021	0.021	0.048
	CPP (pruning)	0.018	0.018	0.018	0.018	0.044
	Fujikawa	0.023	0.023	0.023	0.023	0.048
	Fujikawa (pruning)	0.020	0.020	0.020	0.020	0.042
Global Alt	CPP	0.977	0.977	0.977	0.977	.
	CPP (pruning)	0.874	0.874	0.874	0.874	.
	Fujikawa	0.970	0.970	0.970	0.970	.
	Fujikawa (pruning)	0.948	0.948	0.948	0.948	.
One in the Middle	CPP	0.972	0.972	0.877	0.996	.
	CPP (pruning)	0.874	0.874	0.584	0.979	.
	Fujikawa	0.959	0.959	0.824	0.996	.
	Fujikawa (pruning)	0.947	0.947	0.761	0.993	.
Linear	CPP	0.247	0.566	0.805	0.942	0.247

	CPP (pruning)	0.067	0.379	0.743	0.917	0.067
	Fujikawa	0.236	0.553	0.807	0.944	0.236
	Fujikawa (pruning)	0.162	0.553	0.817	0.938	0.162
Good Nugget	CPP	0.075	0.075	0.075	0.629	0.154
	CPP (pruning)	0.061	0.061	0.061	0.639	0.169
	Fujikawa	0.087	0.087	0.087	0.602	0.178
	Fujikawa (pruning)	0.085	0.085	0.085	0.639	0.181
Bad Nugget	CPP	0.322	0.940	0.940	0.940	0.322
	CPP (pruning)	0.067	0.870	0.870	0.870	0.067
	Fujikawa	0.288	0.936	0.936	0.936	0.288
	Fujikawa (pruning)	0.169	0.938	0.938	0.938	0.169
Half	CPP	0.179	0.179	0.839	0.839	0.278
	CPP (pruning)	0.066	0.066	0.843	0.843	0.128
	Fujikawa	0.176	0.176	0.852	0.852	0.274
	Fujikawa (pruning)	0.159	0.159	0.855	0.855	0.289

Rejection rates and FWERs are shown in Table 12. Results without pruning are again included for better comparison. As could be expected, pruning on average leads to lower power but also lower FWER as pruning prevents the null hypotheses of baskets with a number of responses smaller than c_{pool} from being rejected. Both power loss and lower FWER are more pronounced with the CPP method. In the Global Alternative scenario, pruning reduces the power by 10 percentage points with the CPP method but only by 2 percentage points for Fujikawa's design. In the One in the Middle scenario a similar loss in power is seen. In the Linear scenario, power loss in the second basket is even more dramatic with almost 20 percentage points with the CPP method, but the FWER is also much lower - 7 percent instead of 25 percent. With Fujikawa's design, changes in power are minimal in this scenario but pruning still reduces the FWER by 7 percentage points. In the Good Nugget scenario pruning had little effect for both methods. In the Bad Nugget scenario, the FWER for the CPP method is reduced by 25 percent, but power in the active baskets is lower by 7 percentage points. FWER was also reduced in Fujikawa's design without any consequences regarding power. Finally in the Half scenario, pruning reduces the FWER by 15 percentage points for the CPP method without losing power. With Fujikawa's design, pruning had negligible effects.

To summarise, with the suggested pruning method nonmonotonicity was resolved for all relevant choices of tuning parameters for both the power prior design with CPP weights and

Fujikawa's design. Consequences regarding the performance are, however, quite large. In most scenarios, pruning leads to a lower FWER but also to lower power. Thus if a large inflation of the FWER is a concern, pruning could be applied even when violation of the monotonicity conditions is acceptable.

3.4 R Package baskexact

In this section, the R package baskexact, that implements Fujikawa's design and many power prior design variations is described. As already mentioned, since the posterior of the power prior design is a beta distribution and the posterior parameters are simply calculated by adding weighted sums of the responses and non-responses in the baskets to the prior parameters, posterior probabilities are cheap to calculate. When the number of baskets is small (up to 5), then analytical calculation of typical operating characteristics such as TOERs and power is also feasible. baskexact enables efficient analytical computation of operating characteristics for single-stage and two-stage basket trials with equal sample sizes and equal prior distributions.

3.4.1 Usage

To use the package, at first the user has to create a design object using one of the two setup-functions: `setupOneStageBasket` for a single-stage basket trial or `setupTwoStageBasket` for a basket trial with one interim analysis. For example:

```
> library(baskexact)
> design <- setupOneStageBasket(k = 3, shape1 = 1, shape2 = 1, p0 = 0.2)
```

Here, `k` refers to the number of baskets and `p0` to the response probability under the null hypothesis. `shape1` and `shape2` refer to the prior shape parameters of the beta distribution, s_1 and s_2, respectively. By default these parameters are set to 1. The setup function for the two-stage design has the same arguments.

The idea of creating a design object is that it contains the most basic elements of the design, that are commonly determined in advance and not varied in the planning process. Thus, when calling a function of the package, the design object is always the first argument and

the user does not have to repeatedly specify these elements since they are contained in the object.

The operating characteristics can be calculated with the functions `toer` for TOERs, `pow` for the power, `ecd` for the ECD and `ess` for the expected sample size - the latter only for two-stage trials. With `estim` the mean posterior means and mean squared errors can be calculated. These functions mostly take the same arguments. For example, to calculate the TOER:

```
> toer(
>   design = design,
>   p1 = NULL,
>   n = 20,
>   lambda = 0.99,
>   weight_fun = weights_cpp,
>   weight_params = list(a = 2, b = 2),
>   globalweight_fun = globalweights_fix,
>   globalweight_params = list(w = 0.7),
>   results = "group"
> )


$rejection_probabilities
[1] 0.009493424 0.009493424 0.009493424


$fwer
[1] 0.02232409
```

`p1` is the response probability or vector of response probabilities of the baskets. By default with `p1 = NULL` the TOER is calculated under the global null hypothesis, but it is also possible to calculate the TOER under mixed scenarios as long as at least one basket is truly inactive. `n` is the sample size per basket and `lambda` the probability threshold λ. With `weight_fun`, the type of weight function used to calculate the pairwise weights can be specified. The tuning parameters (passed as a list), that are specific to the function

used to calculate the weights, are set with `weight_params`. The options for the power prior design are `weight_cpp`, `weight_jsd` and `weight_mml` to use weights based on the CPP, the JSD and the MML approach, respectively. Furthermore, `weight_fujikawa` implements Fujikawa's design, which uses the same weights as `weight_jsd` but also shares the prior information. `globalweight_fun` specifies which function to calculate the global weights should be used. Tuning parameters are passed again as a list via the `globalweight_params` argument. By default `globalweight_fun` is set to `NULL` and thus only pairwise weights are used. The options are `globalweights_fix` and `globalweights_diff` to use a fixed global weight or global weights based on the heterogeneity function h, respectively. The global weight function based on the JSD is not implemented since it is computationally too expensive when operating characteristics are computed analytically. Finally, `results = "group"` specifies that not only the FWER but also the basket-wise TOERs are calculated. In the example above, the true response probability is equal to p_0 in all baskets, thus the output under `rejection_probabilities` are the TOERs, but rejection probabilities (thus corresponding to the power) are also calculated in scenarios with some active baskets. The other option is `"fwer"`, which only calculates the FWER.

When the design object refers to a two-stage design, additional arguments are required to specify the details of the interim analysis. For example:

```
> design2 <- setupTwoStageBasket(k = 3, p0 = 0.2)
> toer(
>   design = design2,
>   p1 = NULL,
>   n = 20,
>   n1 = 10,
>   lambda = 0.99,
>   interim_fun = interim_postpred,
>   interim_params = list(prob_futstop = 0.1, prob_effstop = 0.9),
>   weight_fun = weights_cpp,
>   weight_params = list(a = 2, b = 2),
>   globalweight_fun = globalweights_fix,
```

```
>   globalweight_params = list(w = 0.7),
>   results = "group"
> )

$rejection_probabilities
[1] 0.01396859 0.01396859 0.01396859

$fwer
[1] 0.03748156
```

`n1` defines the sample size per basket for the interim analysis. `interim_fun` specifies the type of interim analysis. `interim_postpred` refers to an interim analysis based on the posterior predictive probabilities, as was suggested by Fujikawa et al. (2020). Another option available in baskexact is `interim_posterior` which implements interim decisions based on the posterior probabilities. In `interim_params`, the parameters for the interim analysis are specified. `prob_futstop` is the threshold for a futility stop and `prob_effstop` the threshold for an efficacy stop. Hence, with the values specified in the example above, a basket is stopped for futility if the posterior predictive probability for success is less than 0.1 and for efficacy if this probability is greater than 0.9. See Section 2.4.2 for details about the interim analysis based on the posterior predictive probabilities.

To find the largest probability threshold λ_α that ensures that the FWER is below a certain level, the function `adjust_lambda` can be used. For example:

```
> adjust_lambda(
>   design = design,
>   alpha = 0.05,
>   n = 20,
>   weight_fun = weights_cpp,
>   weight_params = list(a = 2, b = 3),
>   prec_digits = 3
)
```

```
$lambda
[1] 0.974


$toer
[1] 0.04555955
```

The main arguments are again the same (default values for `p1` and `globalweight_fun` are used in this example). The additional parameter in this function is `prec_digits` which specifies the number of decimal places up to which `lambda` is determined. Since the outcome is binary, usually there will be no λ_α for which the FWER is exactly α, but increasing the number of of decimal places of λ might result in a FWER closer to the nominal level. In the example above for instance, if `prec_digits` is increased to 4, then the solution is $\lambda_\alpha = 0.9738$ which gives a FWER of 0.0498, hence much closer to the nominal level of 0.05. However, further increasing `prec_digits` has no additional benefit.

The monotonicity conditions defined in Section 3.3.3 can be checked using the functions `check_mon_within` and `check_mon_between` for the within-trial and the between-trials monotonicity condition, respectively. The main arguments are again the same. When only a single value is specified for every tuning parameter, then if the argument `details` is set to `TRUE` all results that violate the respective monotonicity condition are shown - otherwise the function only returns `TRUE` or `FALSE` to indicate whether the monotonicity condition holds. For illustration, the following function call corresponds to the example in Section 3.3.1:

```
> design <- setupOneStageBasket(k = 4, p0 = 0.15)
> check_mon_within(
>   design = design,
>   n = 20,
>   lambda = 0.99,
>   weight_fun = weights_cpp,
>   weight_params = list(a = 1.5, b = 0.5),
>   details = TRUE
> )
```

```
$Events
[1] 5 5 5 6

$Results
[1] 1 1 1 0
```

Under `$Events` the outcome vector $\boldsymbol{r}$ that violates the within-trial monotonicity rule is shown and under `$Results` it is shown for which of the baskets the null hypothesis is rejected. 1 means that it is rejected and 0 means that it is not.

This function call corresponds to the first example in Section 3.3.2:

```
> check_mon_between(
>   design = design,
>   n = 20,
>   lambda = 0.97,
>   weight_fun = weights_mml
> )

[...]

[[2]]
[[2]]$Events
[,1] [,2] [,3] [,4]
[1,]    0    1    5    6
[2,]    1    3    5    6
[3,]    2    2    5    6
[4,]    2    3    5    6
[5,]    3    3    5    6

[[2]]$Results
[,1] [,2] [,3] [,4]
[1,]    0    0    1    1
```

```
[2,]    0    0    0    0
[3,]    0    0    0    0
[4,]    0    0    0    0
[5,]    0    0    0    0
[...]
```

Only parts of the output of the above function call is shown, since there are more outcome vectors that violate the between-trial monotonicity condition. Under `$Events`, the first row displays the outcome vector for which at least one null hypothesis can be rejected (as seen under `$Results`) and the rows below show the outcomes for which at least one basket has a higher number of responses but no null hypothesis can be rejected.

When a vector of tuning parameters is passed, then a d-dimensional array is returned, where d is the number of tuning parameters for which a vector is passed. For example:

```
> check_mon_between(
>   design = design,
>   n = 20,
>   lambda = 0.99,
>   weight_fun = weights_cpp,
>   weight_params = list(a = 1:3, b = 1:3)
> )


   b
a       1     2     3
  1  TRUE  TRUE  TRUE
  2  TRUE  TRUE FALSE
  3 FALSE FALSE FALSE
```

This also works when a vector is passed to `globalweight_params` when a function for computing global weights is specified.

The function `opt_design` selects the optimal tuning parameters for a design with a specific weight function, where optimisation is performed in the same ways as in the comparison

study in Section 3.2: Based on the specified vectors for the tuning parameters, a grid search is performed, where for each combination of tuning parameter values at first λ is chosen such that the FWER under the global null hypothesis is smaller than or equal to a certain level and then the mean ECD across the specified scenarios is calculated. With the function `get_scenarios` a scenario matrix is created which, based on an alternative response probability, contains $K + 1$ scenarios with an increasing number of active baskets:

```
> get_scenarios(design, p1 = 0.4)
```

```
     0 Active 1 Active 2 Active 3 Active 4 Active
[1,]     0.15     0.15     0.15     0.15      0.4
[2,]     0.15     0.15     0.15     0.40      0.4
[3,]     0.15     0.15     0.40     0.40      0.4
[4,]     0.15     0.40     0.40     0.40      0.4
```

```
> opt_design(
>   design = design,
>   n = 20,
>   alpha = 0.05,
>   weight_fun = weights_cpp,
>   weight_params = list(a = 1, b = 1),
>   scenarios = get_scenarios(design, p1 = 0.4),
>   prec_digits = 3
> )
```

```
  a b Lambda 0 Active 1 Active 2 Active 3 Active 4 Active Mean ECD
1 3 2  0.988 3.923076 3.478154 3.433560 3.522391 3.769971 3.625430
2 2 1  0.985 3.926489 3.482139 3.392680 3.478313 3.833700 3.622664
3 3 3  0.987 3.915711 3.418550 3.333606 3.482182 3.900351 3.610080
4 2 2  0.983 3.907648 3.343371 3.292666 3.484193 3.927709 3.591117
5 3 1  0.989 3.934446 3.544044 3.415599 3.485650 3.490657 3.574079
6 2 3  0.978 3.889992 3.219442 3.099822 3.338827 3.974875 3.504592
```

```
7 1 1  0.977 3.904800 3.264166 3.097785 3.257325 3.976660 3.500147
8 1 2  0.973 3.880758 3.100604 2.871991 3.172228 3.991116 3.403339
9 1 3  0.972 3.877145 2.974921 2.710836 3.110448 3.994395 3.333549
```

The argument `prec_digits` is used as in the function `adjust_lamda`. In the output, `Lambda` is the selected probability threshold that protects the one-sided FWER under the global null hypothesis. The results are sorted decreasingly by the values in the column `Mean ECD`.

3.4.2 Implementation Details and Computational Efficiency

Since all operating characteristics are calculated analytically with baskexact, one of the main design principles of the R package was computational efficiency to have acceptable computation times. One important step to achieve that is that weights used for the pairwise information sharing are only computed once in the beginning. Specifically, the functions that can be passed to the argument `weight_fun`, such as `weight_cpp` create a matrix with all possible pairwise sharing weights that can occur. In a single-stage design, there are only $(n+1)^2$ possible weights for pairwise sharing (the $+1$ is necessary since 0 is also a possible outcome) and in a two-stage design, $(n+n_1+2)^2$ weights are possible. (Note that with some weight functions not all $(n+1)^2$ weights in the matrix may be unique, as e.g. the CPP weights only depend on the rate differences between baskets. However, the full matrix is always computed.) Additionally to all possible weights for pairwise sharing, the matrix that is returned by the weight functions is assigned a class, either `pp`, for the power prior design, or `fujikawa`, for Fujikawa's design. The class determines whether only the observed information or also the prior parameters are shared in the functions that compute the posterior parameters. Computation of the posterior distribution parameters after sharing is performed by an internal function called `beta_borrow`.

To further reduce computational time, some internal functions were written in C++ using the R packages Rcpp (Eddelbuettel et al., 2023a) and RcppArmadillo (Eddelbuettel et al., 2023b). Furthermore, some of the functions are parallelised using the doFuture package (Bengtsson, 2023, 2021). Other than with older R packages that facilitate parallelisation, with doFuture only a single line of code is necessary for the user of the package to run code that was written using doFuture (e.g. for-loops) in parallel. Various backends for parallelisation are possible,

e.g. depending on the operating system. On a standard Windows personal computer, the following line can be run:

```
> plan(multisession, worker = 4)
```

This makes all parallelised code run in parallel on 4 cores.

baskexact was implemented using R's S4 and S3 class systems. The objects created by the functions `setupOneStageBasket` and `setupTwoStageBasket` are S4 objects and many functions available to the user utilise the S4 class system. More specifically, most functions are internally available in two varieties - one for single-stage designs and one for two-stage designs. The correct version is chosen based on the class of the design object that is passed to it, which is called method dispatch. All functions for the computation of operating characteristics as well as `adjust_lambda` and `opt_design` are available for single-stage and two-stage designs. The functions that check the monotonicity conditions are only available for single-stage designs since the monotonicity conditions as proposed in Section 3.3.3 are not well defined for two-stage trials.

Internally, S3 classes are used. As mentioned above, the weight matrices created by the weight functions such as `weight_cpp` are either of class `pp` or `fujikawa`. These are S3 classes, and the internal function `beta_borrow` also exists in two variants to handle the computation of the posterior distribution parameters for both classes.

baskexact was also written to be easily extendable. Specifically, new methods to compute pairwise or global sharing weights can be easily implemented. For the pairwise weights, a new set of functions has to be written using R's `setMethod` function for writing S4 functions. Due to the object oriented programming approach, it is necessary to create a so called generic function, which handles the method dispatch, and two functions for single-stage and two-stage designs, respectively. The functions must return a weight matrix in the correct format and assign either class `pp` or `fujikawa` to it. The new weight functions can then be passed to the `weight_fun` argument of all functions as demonstrated in the previous section. Since the global weights do not differentiate between single-stage and two-stage design, a standard R function can be created that returns a single number (the weight). This function can then be

passed to the `globalweight_fun` argument. The package includes a vignette which explains in detail how new weight functions can be implemented.

3.4.3 Computation Times

Since operating characteristics are calculated analytically with baskexact, the computation times heavily depend on the parameters of the design, most importantly the number of baskets but also e.g. sample size and, in a two-stage design, stopping rules. Table 13 shows computation times of FWER and basket-wise power for a single-stage and a two-stage power prior design with CPP weights, with a sample size of $n = 20$ per basket and 3, 4 and 5 baskets. For the two-stage design, an interim analysis is performed after $n_1 = 10$ observations per basket, based on the posterior predictive probability. Baskets are stopped for futility if this probability is below 0.1 and for efficacy if it is above 0.9. Computations were done on a personal computer with an Intel Core i7-13700 processor with 2.1 GHz. Note that the two functions `toer` and `pow`, which are used in these calculations, do not utilise parallelisation, since for a small number of baskets the computation times are already low and parallelisation introduces computational overhead which would actually increase the computation times in some cases. Times were measured using the R package microbenchmark (Mersmann, 2023). The median computation time of 5 function calls is reported, except for power in the two-stage design which was only called once.

Table 13: *Computation times of baskexact with a single-stage and a two-stage design*

	3 Baskets	4 Baskets	5 Baskets
FWER (Single-Stage)	62 msec	287 msec	1.5 sec
Power (Single-Stage)	208 msec	9.9 sec	1.9 min
FWER (Two-Stage)	524 msec	9.2 sec	4.0 min
Power (Two-Stage)	3.8 sec	2.1 min	65.6 min

With 3 and 4 baskets, computation times are low to moderate. Computations in the single-stage design take at most 10 seconds for the calculation of basket-wise power and at most around 2 minutes for the basket-wise power of a two-stage design. With 5 baskets, computations with the single-stage design are still moderate, but computation of basket-wise power in the two-stage design already exceeds one hour.

Note that FWER is computed much faster than the power due to a faster implementation of experiment-wise operating characteristics - computation times of basket-wise TOERs would be similar to those of basket-wise power.

3.4.4 Validation

A further design principle of baskexact was to ensure (to a certain degree) that the provided output is correct. For that, the R package testthat (Wickham, 2023, 2011) was utilised. With testthat, tests can be written that check whether a certain function call produces the expected result. These tests can be run manually with the function `test`, but are also run when a package is submitted to CRAN. Every line of baskexact is covered by at least one test and in total more than 300 tests were written. For instance, many of the results from Fujikawa et al. (2020) were used as a basis for tests. Other tests are based on internal consistency. For example, it is tested whether the value for λ and the resulting FWER computed by `adjust_lambda` can be reproduced when `toer` is called, and whether the nominal level is in fact exceeded when $10^{-\texttt{prec\_digits}}$ is subtracted from λ.

Furthermore, several functions were implemented in two different ways. The functions that compute the operating characteristics are generally based on creating a matrix of all possible outcomes and then analysing all relevant outcomes using `apply`. Additionally, internal validation functions based on for-loops were written, that are usually slower but easier to follow and do not use any computational shortcuts. Several tests compare the output of the two versions of a function that compute the same result.

3.5 R Package basksim

The basksim package facilitates simulation based computation of operating characteristics for single-stage basket trial designs with equal sample sizes. The supported designs include methods also implemented in baskexact, bhmbasket and bmabasket, in order to provide a unified syntax for comparisons.

3.5.1 Usage

To use basksim, as in baskexact, a design object has to be created by the user with a setup function. In basksim, the type of design object is not determined by the number of stages but by the design that shall be used for the analysis. For example, to setup a design object for the MML-Global method:

```
> library(basksim)
> design <- setup_mmlglobal(k = 3, p0 = 0.2, shape1 = 1, shape2 = 1)
```

All available setup functions are shown in Table 14.

Table 14: *Setup functions available in the basksim package*

Function	Method
`setup_bhm`	BHM
`setup_bma`	BMA
`setup_cpp`	CPP
`setup_cppglobal`	CPP-Global
`setup_exnex`	EXNEX
`setup_fujikawa`	Fujikawa
`setup_jsdglobal`	JSD-Global
`setup_mml`	MML
`setup_mmlglobal`	MML-Global

The main function for the user to calculate the operating characteristics of a design given certain tuning parameters is `get_details`, which returns the rejection probabilities for each basket, the FWER, the posterior means, mean squared error, the mean lower and upper limits of the credibility intervals and the ECD. For example:

```
> get_details(
>   design = design,
>   n = 20,
>   p1 = c(0.2, 0.5, 0.5),
>   lambda = 0.95,
>   level = 0.95,
>   iter = 10000
```

```
> )

$Rejection_Probabilities
[1] 0.2575 0.9593 0.9587

$FWER
[1] 0.2575

$Mean
[1] 0.2566678 0.4855678 0.4828403

$MSE
[1] 0.01181304 0.01025248 0.01015605

$Lower_CL
[1] 0.1179411 0.3375865 0.3352567

$Upper_CL
[1] 0.4065638 0.6336367 0.6306097

$ECD
[1] 2.6605
```

The arguments `n`, `p1` and `level` have the same meaning as in the baskexact package. Additionally, the `level` argument here specifies the level of the credibility intervals and `iter` sets the number of randomly created data sets that are used to estimate the operating characteristics. Since a binary endpoint is investigated, a single data set is simply a vector of length K which contains the number of observed responses per basket. In the example above, 10,000 data sets are created, but if a certain matrix of data sets should be used instead (e.g. for better reproducibility), a data matrix can be passed to `get_details` via the argument `data`. Random data can be generated with the function `get_data`.

Another main function is `get_results`, which has the same arguments as `get_details` and returns an `iter`$\times$`k` matrix of 0s and 1s, where a 0 represents an unrejected null hypothesis and a 1 a rejected null hypothesis. Two further important functions for users are `opt_design` and `adjust_lambda` which work in the same way as in baskexact and are therefore not illustrated here. Further available functions are `toer` and `ecd` to calculate the FWER and ECD of a design, respectively. However, these are mainly used internally and provide no benefit as compared to using `get_details`.

3.5.2 Implementation Details

Since basksim computes all results based on simulations, the implementation is much less complex than that of baskexact. This is why with `get_details` a function is available that returns several operating characteristics with just one function call. For the power prior based methods and Fujikawa's design, when `get_details` is called, a for-loop evaluates all `iter` data sets by at first computing the posterior parameters after information sharing and based on these estimates all operating characteristics. Since the computationally most expensive part is to calculate the posterior parameters, calculating all operating characteristics in one function adds little additional computation time and complexity to the implementation.

For the implementation of the BMA, BHM and EXNEX designs, functions in basksim are mostly wrappers for functions of the bhmbasket and bmabasket package. For example, in the implementation of `get_details` for the BMA design, the function `bma` of bmabasket is used, which returns the posterior probabilities that the response probability is greater than p_0 and the posterior means for a given vector of responses. The rejection probabilities, FWER, mean posterior means and mean squared errors are then calculated based on the output of `bma`. Since no function to compute credibility intervals is available in bmabasket, BMA is the only design for which `get_details` does not return mean credibility interval limits.

basksim also utilises object-oriented programming. The setup functions create S3 class objects and `get_details` and `get_results` have a different implementation for each class. All other functions, however, are standard functions without class specific implementations. They still work for all methods, since internally they call the `get_results` function. Therefore, it is still

relatively easy to implement new methods in basksim. Only a new version for `get_details` and `get_results` must be provided for the new design.

As for baskexact, every line of code is tested with at least one unit test (more than 200 unit tests in total) and the doFuture package is utilised to provide shorter computation times with parallelised code.

Chapter 4

Discussion

In this chapter, the results of the thesis and its contribution to the basket trial literature are discussed. Furthermore, limitations and topics for further research are given. The discussion is split into two parts corresponding to the two main topics of the thesis. In Section 4.1, the power prior design and the results of the comparison study are covered. Section 4.2 discusses the nonmonotonic decisions in basket trials and the proposed monotonicity conditions. The chapter ends with a conclusion.

4.1 Power Prior Design and Comparison Study

4.1.1 Discussion and Contributions to Research

In this thesis, a new basket trial design based on power priors was presented which was shown to be closely related to a design by Fujikawa et al. (2020). Different ways to calculate the weights that determine the amount of shared information were suggested, which were mainly adapted from methods in the power prior literature. Additionally, the idea of using global weights was introduced which quantify the heterogeneity across all baskets. In an extensive comparison study, the variations of the power prior design were compared to Fujikawa's design and three other Bayesian basket trial designs.

Fujikawa's design has not been systematically investigated in a comparison study before. While Fujikawa et al. (2020) provide some simulation results, they only compared the per-

formance of the design with that of a Bayesian hierarchical model (Thall et al., 2003) and a basket-wise analysis with $K = 3$ baskets and $n = 24$ per basket in a two-stage design in four scenarios with an increasing number of active baskets and a common response probability for active baskets. The authors argued that their design results in higher power and lower excepted sample size, but the probability threshold λ was not tuned such that the FWER was protected at a certain level and the tuning parameters were not selected based on any optimality criteria.

The results of the comparison study in Section 3.2.3 show that when the tuning parameters are chosen to be optimal in terms of the mean ECD and λ is tuned such that the FWER under the global null hypothesis is at most 5%, then the results of Fujikawa's design and other Bayesian basket trial designs is very similar. Small improvements could be achieved by calculating the sharing weights in different ways in the power prior design, but adding a global weight based on the observed overall heterogeneity had no additional beneficial effect.

The results of the comparison study are in line with the results of Broglio et al. (2022). In their simulation study, the authors compared the BHM and the EXNEX design, as well as the multisource exchangeability model (MEM) design and the ROAR design (named after the trial in which the model was used). In the MEM design (Hobbs and Landin, 2018) information between baskets is shared using a very similar approach as in the BMA design. In the ROAR design (Subbiah et al., 2020), baskets are at first clustered, where the number of clusters is determined from the data, and then in each cluster a Bayesian hierarchical model is calculated. Broglio et al. (2022) investigated the performance of these four designs in $K = 8$ baskets with $n = 20$ per basket, and also found very minor differences when the designs were optimised based on the mean ECD across 6 scenarios (Global Null, Global Alternative, Good Nugget, Bad Nugget, Half and Linear, but with different response probabilities than used in Section 3.2 due to the higher number of baskets). They report that this was also the case, when the target for optimisation was not the mean ECD but a utility function that put greater penality on type 1 errors than on type 2 errors, and when the FWER under the global null hypothesis was not controlled. The authors conclude that increasing the complexity of the designs does not lead to a relevant benefit in performance.

Note that a preprint was published on arXiv on 23 December 2023 by Zhou et al. (2023), also suggesting the analysis of basket trials using power priors. I became aware of this preprint on 20 March 2024 shortly before finalising this thesis. Those parts of the results presented in Section 3.1 and 3.2 of my thesis, for which there is some overlapping with Zhou et al. (2023) regarding the power prior design, were published in an arXiv preprint, the first version of which was uploaded already on 13 September 2023 (see also Chapter 8)."

The main benefit of the power prior design - and thus in a broader sense also of Fujikawa's design - besides easy interpretability is computational speed, as calculation of posterior probabilities does not require MCMC sampling and operating characteristics for many weights can be computed analytically. Another advantage of the computational efficiency of the design is the ability to identify nonmonotonic outcomes.

One characteristic of the power prior design results from the fact that the design is not fully Bayesian. There is no straightforward way to incorporate prior information about the similarity of subgroups into the design. The weights are calculated based on the observed data only and there is no prior distribution involved. In the BMA design, for example, model probabilities are updated based on the observed data and could then be used as the prior probabilities in a further trial. In the power prior design, it would be, however, possible to use different tuning parameters for different weights corresponding to the information that is shared between different pairs of baskets. For example, if it is known a priori that two baskets respond similarly to the investigated treatment then tuning parameters for the pairwise calculation of these two baskets could be selected such that more information is shared even when the observed response rates differ.

4.1.2 Limitations and Directions for Future Research

Due to the high number of design elements in a basket trial, the comparison study necessarily has some limitations. First, sample sizes were assumed to be equal in all baskets which is usually not the case in actual basket trials. When there is little fluctuation in the sample sizes of the baskets, the consequences on the performance of the power prior design and other methods are expected to be minor. In real basket trials, however, the sample size differences in the baskets can be tremendous. In the ROAR trial (Subbiah et al., 2023), for example, the

sample sizes in the 8 baskets ranged from 1 to 55. Hyman et al. (2015) report the results of a basket trial (also mentioned in Section 1.1) in which 7 baskets with sample sizes between 5 and 37 were analysed.

When sample size differences are that large and sharing is not adapted to these differences, severe bias in the small baskets is possible. While Neuenschwander et al. (2016) and Berry et al. (2013) consider a scenario with unequal sample sizes, they do not specifically discuss how the design may be adapted to accommodate these sample size differences. In the context of borrowing from historical data, Ollier et al. (2020) proposed a power prior method that limits the amount of information that is shared from the historical data through an additional parameter when the sample size of the historical study has a larger sample size than that of the current study. This could also be applied to basket trials. A systematic examination of how the power prior design and other basket trial designs perform under different sample size scenarios is an important and extensive topic for further research as many different sample size configurations may appear in practice and could be considered.

A further limitation is that a setting with four baskets was assumed for the comparison study and results may thus differ for trials with a different number of baskets. While it seems plausible that the results are similar for a trial with e.g 3 or 5 baskets, this is less clear when the number of baskets is larger. When in these settings the amount of information sharing is not limited, sharing between a large number of baskets could also lead to high bias.

Furthermore, due to the focus on the information sharing component, a single-stage design was used. Several research questions concerning the application of interim analyses in basket trials are still unanswered, such as the optimal number, timing and type of interim analyses. Interim analyses are used in many basket trial simulations and their benefit with respect to expected sample size is apparent, but there are no systematic comparison studies investigating the influence of different types of interim analyses.

4.2 Nonmonotonic Decisions and Monotonicity Conditions

4.2.1 Discussion and Contributions to Research

Information sharing between subgroups is the defining design element of a basket trial from a statistical perspective. While in terms of the operating characteristics information sharing has clear benefits, it was shown in this thesis that data dependent information sharing can also lead to decisions that are not monotonically increasing in the number of observed responses. The issue has been mentioned by Kopp-Schneider et al. (2020) in the context of borrowing from a single historical study with a binary endpoint. The authors showed that an extreme borrowing approach that only utilises the historical data when the observed response rates of the historical and the current data coincide exactly can lead to a nonconnected rejection region.

In basket trials, two types of nonmonotonicity - within-trial and between-trial nonmonotonicity - were identified. Using different examples it was argued that in some cases nonmonotonicity could be appropriate, for example when prior information is available that suggests that some baskets respond similarly to the treatment under investigation. Thus, other than in the single-arm trial example given in Kopp-Schneider et al. (2020), where it is clearly unacceptable, the relevance of monotonicity is ambiguous in the context of basket trials. Nevertheless, with the monotonicity conditions proposed in this thesis researchers can explore which nonmonotonic events may occur and can then decide whether the violations are acceptable. Due to the computational efficiency of the power prior design, nonmonotonic events can be detected easily for a moderate number of baskets using the baskexact R package.

To prevent nonmonotonicity, a pruning method was suggested in which baskets that do not achieve a certain number of responses are pruned, i.e. excluded from the information sharing. For the power prior design with CPP weights and Fujikawa's design, it was shown that events that violate the monotonicity conditions can occur for a range of different tuning parameters. With Fujikawa's design, the between-trial monotonicity condition was violated for all investigated tuning parameter values when $K \geqslant 5$, but the within-trial condition held in most cases. With the CPP method, there were still many parameter values for which the

between-trial condition was not violated, but there were more cases in which the within-trial condition did not hold. When pruning was applied, nonmonotonicity was resolved for most tuning parameter values. When pruning is used, it has to be considered, however, that this also affects the operating characteristics and leads to lower power but also lower inflation of the FWER.

While for the investigation only the power prior design with CPP weights and Fujikawa's design were used, violation of the monotonicity condition is also expected to occur with other basket trial designs in which information is shared based on observed similarity. For the BMA design, events that violate the between-trial monotonicity condition can also be found. Identifying nonmonotonic events in designs for which posterior probabilities cannot be computed analytically is difficult, as often posterior probabilities that are exact up to the second or third decimal place are necessary to decide whether nonmonotonicity is present. This is computationally expensive to achieve when posterior probabilities are based on MCMC sampling, such as in the BHM and EXNEX design. Within-trial nonmonotonicity may, however, be resolved entirely by using designs such as the BHM, which model the transformed response probabilities of all baskets using a single distribution.

4.2.2 Limitations and Directions for Future Research

As in the entire thesis, equal sample sizes and equal prior distributions in all baskets were assumed. Without this assumption the definition of the monotonicity conditions in Section 3.3.3 based on the absolute number of responses would not make sense and therefore would have to be modified. This is also relevant when interim analyses are conducted. When stopping early is only possible for futility, then no modification is necessary, but when baskets can also be stopped early for efficacy, the same issue arises, as null hypotheses in baskets with different sample sizes then have the possibility to be rejected.

More general monotonicity conditions could be based on the basket-wise posterior probabilities, i.e. $\mathbb{P}(p_k > p_0|r_k)$ instead of r_k. For equal sample sizes this would clearly lead to the same results. The definition of the critical pooled value c_{pool} is, however, not easily generalisable. A threshold for the basket-wise posterior probabilities could be used but, for example, a grid search would be necessary to determine the cut-off that ensures that there is no violation

of the monotonicity conditions. More research is necessary to explore nonmonotonicity and possible solutions in basket trials with unequal sample sizes.

4.3 Conclusion

In this thesis, the power prior design for the analysis of basket trials was proposed, which can be seen as an extension of Fujikawa's design. In a comparison study, Fujikawa's design and the power prior design were shown to have similar operating characteristics to other Bayesian basket trial designs in a single-stage trial with equal sample sizes. While improvements of the power prior design compared to Fujikawa's design were very small, the different variations of the power prior design allow to more flexibly tune the amount of information that is shared between baskets based on the results observed in the trial.

One of the main advantages of the power prior design is that the sharing mechanism is driven by weights that are easy to interpret as they represent the percentage of data that are used from another basket in the computation of the posterior probabilities. A further benefit is computational efficiency. Posterior probabilities can always be calculated analytically and in some cases even analytical computation of TOER, power and other operating characteristics is achievable. This also enables identification on nonmonotonic events.

While further research is necessary to investigate the performance of the power prior design in settings with unequal sample sizes and interim analyses, the results of this thesis suggest that the power prior design is suitable for the analysis of basket trials.

Chapter 5

Summary

Basket trials are a new type of clinical trial in which a treatment is investigated in several subgroups. They are often used in uncontrolled oncology trials with a binary endpoint such as tumour response. The subgroups, for example, comprise patients with different tumour locations but all patients in the trial share a common genetic feature. Several designs for the analysis of such trials were proposed in the literature. The main element of basket trial designs is information sharing between subgroups depending on the observed similarity. Mostly Bayesian methods have been proposed for that. For example, in Fujikawa's design information is shared based on the pairwise similarity between the individual posterior distributions of the subgroups.

The main objective of this thesis is to extend and improve Fujikawa's design and to compare the performance of the original and the modified design to that of other Bayesian basket trial designs.

It is shown that the sharing mechanism in Fujikawa's design is closely related to power priors, which were originally proposed to borrow strength from historical data. The only difference is that the proposed basket trial design based on power priors only shares the data observed in the trial while Fujikawa's design also shares prior information.

Using this connection, different methods for computing the sharing weights from the power prior literature are adapted to basket trials. While in Fujikawa's design the amount of infor-

mation that is shared between subgroups only depends on their pairwise similarity, approaches that additionally consider the overall heterogeneity were also explored.

In a comparison study, it is demonstrated that the design based on power priors performs similarly to Fujikawa's design and other Bayesian basket trial designs in terms of the expected number of correct decisions and rejection probabilities across a range of different scenarios. The power prior design leads to minimal improvements compared to Fujikawa's design. Considering the overall heterogeneity had, however, no additional benefits. However, with the different power prior variants better fine tuning of the information sharing is possible.

It is also shown that information sharing in basket trials can lead to a number of rejected null hypotheses that is not monotonically increasing in the number of observed events. Two types of nonmonotonicity are identified and monotonicity conditions are proposed. Results that violate these conditions can occur in Fujikawa's design and the power prior design. A pruning strategy is suggested that helps to prevent nonmonotonicty in many cases but also has relevant influence on the operating characteristics.

Two R packages, baskexact and basksim, in which the power prior design is implemented, were developed. A benefit of the power prior design is that it is computationally very cheap such that posterior probabilities can be calculated analytically and even analytical computation of operating characteristics is feasible in some cases.

The finding that, regardless of their complexity, different basket trial designs perform similarly, is in line with the existing literature. While further research is necessary to investigate the power prior design in settings with different sample sizes per basket and with interim assessments, the design is attractive as the sharing can be flexibly tuned and as it is computationally cheaper than other Bayesian basket trial designs.

Chapter 6

Zusammenfassung

Basket-Studien sind ein neuer Typ klinischer Studien, in denen eine Behandlung in verschiedenen Subgruppen untersucht wird. Sie werden häufig in unkontrollierten onkologischen Studien mit einem binären Endpunkt wie Tumoransprechen eingesetzt. Die Subgruppen setzen sich beispielsweise aus Patientinnen und Patienten mit verschiedenen Tumorlokalisationen zusammen, die ein gemeinsames genetisches Merkmal aufweisen. Für die Analyse solcher Studien wurden verschiedene Designs in der Literatur vorgeschlagen. Das Kernelement von Designs für Basket-Studien ist, dass Information zwischen den Subgruppen, basierend auf der beobachteten Ähnlichkeit, geteilt wird. Dafür wurden hauptsächliche Bayesianische Methoden vorgeschlagen. In Fujikawas Design beispielsweise wird Information basierend auf der paarweisen Ähnlichkeit der individuellen A-posteriori-Verteilungen der Subgruppen geteilt.

Das Hauptziel dieser Dissertation ist es, Fujikawas Design zu erweitern und zu verbessern und die Performance des ursprünglichen und des modifizierten Designs mit der anderer Bayesianischer Basket-Studiendesigns zu vergleichen.

Es wird gezeigt, dass der Mechanismus zum Teilen von Information in Fujikawas Design eng verwandt mit der Power-Prior-Methode ist, die ursprünglich vorgeschlagen wurde, um historische Daten zu nutzen. Der einzige Unterschied ist, dass das vorgeschlagene Basket-Studiendesign basierend auf der Power-Prior-Methode nur die Daten teilt, die in der Studie beobachtet wurden, wohingegen in Fujikawas Design auch A-priori-Information geteilt wird.

Basierend auf dieser Verbindung werden verschiedene Methoden zum Berechnen der Gewichte, die die Menge an geteilter Information bestimmen, aus der Power-Prior-Literatur für Basket-Studien adaptiert. Während in Fujikawas Design die Menge an Information, die zwischen den Subgruppen geteilt wird, nur von deren paarweiser Ähnlichkeit abhängt, werden auch Ansätze untersucht, die zusätzlich die Heterogenität aller Subgruppen berücksichtigen.

In einer Vergleichsstudie wird gezeigt, dass das Power-Prior-Design ähnliche Ergebnisse wie Fujikawas Design und andere Bayesianische Designs bezüglich der erwarteten Anzahl an korrekten Entscheidungen und Verwerfungswahrscheinlichkeiten in einer Reihe von verschiedenen Szenarien erzielt. Das Power-Prior-Design führt zu einer minimalen Verbesserung gegenüber Fujikawas Design, die Gesamtheterogenität zu berücksichtigen bringt jedoch keinen zusätzlichen Nutzen. Allerdings erlauben die verschiedenen Varianten des Power-Prior-Designs eine bessere Feinabstimmung der Menge an geteilten Daten.

Es wird außerdem gezeigt, dass das Teilen von Information in Basket-Studien zu einer Anzahl von verworfenen Nullhypothesen führen kann, die nicht monoton steigend in der Anzahl an beobachteten Ereignissen ist. Zwei Arten von Nicht-Monotonie werden identifiziert und Monotonie-Bedingungen vorgeschlagen. Ergebnisse, die diese Bedingungen verletzen, können mit Fujikawas Design und dem Power-Prior-Design auftreten. Ein Verfahren zum Beschneiden von Subgruppen wird vorgeschlagen. Dieses verhindert Nicht-Monotonie in vielen Fällen, hat aber auch relevanten Einfluss auf die Operationscharakteristiken.

Zwei R-Pakete, baskexact und basksim, in denen das Power-Prior-Design implementiert ist, wurden entwickelt. Ein Vorteil des Power-Prior-Designs ist die schnelle Berechenbarkeit. So können A-posteriori-Wahrscheinlichkeiten analytisch berechnet werden und sogar die analytische Berechnung von Operationscharakteristiken ist in manchen Fällen möglich.

Das Ergebnis, dass verschiedene Designs für Basket-Studien, unabhängig von ihrer Komplexizät, eine ähnliche Performance aufweisen, ist in Einklang mit der existierenden Literatur. Auch wenn weitere Forschung notwendig ist, um das Power-Prior-Design in Szenarien mit ungleichen Fallzahlen in den Subgruppen oder mit Zwischenauswertungen zu untersuchen, ist das Design attraktiv, da das Teilen von Information flexibel angepasst werden kann und es weniger rechenintensiv als andere Bayesianische Designs für Basket-Studien ist.

Chapter 7

Reference List

Alt, E. (2022). **bmabasket: Bayesian model averaging for basket trials**. R package version 0.1.2, `https://CRAN.R-project.org/package=bmabasket`.

Bengtsson, H. (2021). **A unifying framework for parallel and distributed processing in R using futures**. The R Journal, 13(2), 273–291, doi: 10.32614/RJ-2021-048.

Bengtsson, H. (2023). **dofuture: Use foreach to parallelize via the future framework**. R package version 1.0.0, `https://CRAN.R-project.org/package=doFuture`.

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, doi: 10.1002/pst.2088.

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. M. (2000). **Bayesian theory**. John Wiley & Sons, Chichester New York Weinheim.

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). **Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology**

clinical trials. Clinical Trials, 10(5), 720–734, doi: 10.1177/1740774513497539.

Bolstad, W. M. (2007). **Bayesian statistics**, 2nd edition. John Wiley & Sons, Hoboken.

Broglio, K. R., Zhang, F., Yu, B., Marshall, J., Wang, F., Bennett, M., and Viele, K. (2022). **A comparison of different approaches to bayesian hierarchical models in a basket trial to evaluate the benefits of increasing complexity**. Statistics in Biopharmaceutical Research, 14(3), 324–333, doi: 10.1080/19466315.2021.2008484.

Chen, C., Li, X., Yuan, S., Antonijevic, Z., Kalamegham, R., and Beckman, R. A. (2016). **Statistical design and considerations of a phase 3 basket trial for simultaneous investigation of multiple tumor types in one study**. Statistics in Biopharmaceutical Research, 8(3), 248–257, doi: 10.1080/19466315.2016.1193044.

Chen, C., Zhou, H., Li, W., and Beckman, R. A. (2021). **How many cohorts should be considered in an exploratory master protocol?** Statistics in Biopharmaceutical Research, 13(3), 280–285, doi: 10.1080/19466315.2020.1841022.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). **Bayesian ideas and data analysis: an introduction for scientists and statisticians**. CRC press, Boca Raton.

Cover, T. M. and Thomas, J. A. (2006). **Elements of information theory**, 2nd edition. John Wiley & Sons, Hoboken.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B., and Gönen, M. (2017). **An efficient basket trial design**. Statistics in Medicine, 36(10), 1568–1579, doi: 10.1002/sim.7227.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Ucar, I., Bates, D., and Chambers, J. (2023a). **Rcpp: Seamless R and C++ integration**. R package version 1.0.11, `https://CRAN.R-project.org/package=Rcpp`.

Eddelbuettel, D., Francois, R., Bates, D., Ni, B., and Sanderson, C. (2023b). **RcppArmadillo: 'Rcpp' integration for the 'Armadillo' templated linear algebra library**. R package version 0.12.6.6.0, `https://CRAN.R-project.org/package=RcppArmadillo`.

Freidlin, B. and Korn, E. L. (2013). **Borrowing information across subgroups in phase II trials: is it useful?** Clinical Cancer Research, 19(6), 1326–1334, doi: 10.1158/1078-0432.CCR-12-1223.

Fridlyand, J., Simon, R. M., Walrath, J. C., Roach, N., Buller, R., Schenkein, D. P., Flaherty, K. T., Allen, J. D., Sigal, E. V., and Scher, H. I. (2013). **Considerations for the successful co-development of targeted cancer therapies and companion diagnostics**. Nat Rev Drug Discov, 12(10), 743–755, doi: 10.1038/nrd4101.

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, doi: 10.1002/bimj.201800404.

Gaujoux, R. (2023). **doRNG: Generic reproducible parallel backend for 'foreach' loops**. R package version 1.8.6, `https://CRAN.R-project.org/package=doRNG`.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). **Bayesian data analysis**, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Gravestock, I. and Held, L. (2017). **Adaptive power priors with empirical bayes for clinical trials**. Pharmaceutical Statistics, 16(5), 349–360, doi: 10.1002/pst.1814.

Gravestock, I. and Held, L. (2019). **Power priors based on multiple historical studies for binary outcomes**. Biometrical Journal, 61(5), 1201–1218, doi: 10.1002/bimj.201700246.

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, doi: 10.1016/j.conctc.2018.08.009.

Hobbs, B. P. and Landin, R. (2018). **Bayesian basket trial design with exchangeability monitoring**. Statistics in Medicine, 37(25), 3557–3572, doi: 10.1002/sim.7893.

Hobbs, B. P., Pestana, R. C., Zabor, E. C., Kaizer, A. M., and Hong, D. S. (2022). **Basket trials: review of current practice and innovations for future trials**. J Clin

Oncol, 40(30), 3520–3528, doi: 10.1200/JCO.21.02285.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). **Bayesian model averaging: A tutorial**. Statistical Science, 14(4), 382–417, doi: 10.1214/ss/1009212519.

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., et al. (2015). **Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations**. New Engl J Med, 373(8), 726–736, doi: 10.1056/NEJMoa1502309.

Ibrahim, J. G. and Chen, M.-H. (2000). **Power prior distributions for regression models**. Statistical Science, 15(1), 46–60, doi: 10.1214/ss/1009212673.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). **The power prior: theory and applications**. Statistics in Medicine, 34(28), 3724–3749, doi: 10.1002/sim.6728.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). **On optimality properties of the power prior**. Journal of the American Statistical Association, 98(461), 204–213, doi: 10.1198/016214503388619229.

Jørgensen, J. T. (2019). **A paradigm shift in biomarker guided oncology drug development**. Ann Transl Med, 7(7), 148, doi: 10.21037/atm.2019.03.36.

Kaizer, A. M., Koopmeiners, J. S., Chen, N., and Hobbs, B. P. (2021). **Statistical design considerations for trials that study multiple indications**. Statistical Methods in Medical Research, 30(3), 785–798, doi: 10.1177/0962280220975187.

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, doi: 10.1002/bimj.201800395.

Krajewska, M. and Rauch, G. (2021). **A new basket trial design based on clustering of homogeneous subpopulations**. Journal of Biopharmaceutical Statistics, 31(4), 425–447, doi: 10.1080/10543406.2021.1897993.

Kruschke, J. (2015). **Doing bayesian data analysis: A tutorial with R, JAGS, and Stan**, 2nd edition. Academic Press, London San Diego Waltham.

Lin, J. (1991). **Divergence measures based on the shannon entropy**. IEEE Transactions on Information theory, 37(1), 145–151, doi: 10.1109/18.61115.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). **The BUGS book: A practical introduction to bayesian analysis**. CRC press, Boca Raton London New York.

Mersmann, O. (2023). **microbenchmark: Accurate timing functions**. R package version 1.4.10, `https://CRAN.R-project.org/package=microbenchmark`.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). **A note on the power prior**. Statistics in Medicine, 28(28), 3562–3566, doi: 10.1002/sim.3722.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016). **Robust exchangeability designs for early phase clinical trials with multiple strata**. Pharmaceutical statistics, 15(2), 123–134, doi: 10.1002/pst.1730.

Nielsen, F. (2021). **On a variational definition for the jensen-shannon symmetrization of distances based on the information radius**. Entropy, 23(4), 464, doi: 10.3390/e23040464.

Ntzoufras, I. (2009). **Bayesian modeling using WinBUGS**. John Wiley & Sons, Hoboken.

Ollier, A., Morita, S., Ursino, M., and Zohar, S. (2020). **An adaptive power prior for sequential clinical trials–application to bridging studies**. Statistical Methods in Medical Research, 29(8), 2282–2294, doi: 10.1177/0962280219886609.

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, doi: doi.org/10.1111/rssc.12204.

Pohl, M., Krisam, J., and Kieser, M. (2021). **Categories, components, and techniques in a modular construction of basket trials for application and further research**. Biometrical Journal, 63(6), 1159–1184, doi: 10.1002/bimj.202000314.

Psioda, M. A. (2023). Personal communication.

Psioda, M. A., Xu, J., Jiang, Q., Ke, C., Yang, Z., and Ibrahim, J. G. (2021). **Bayesian adaptive basket trial design using model averaging**. Biostatistics, 22(1), 19–34, doi: 10.1093/biostatistics/kxz014.

Redig, A. J. and Jänne, P. A. (2015). **Basket trials and the evolution of clinical trial design in an era of genomic medicine**. J Clin Oncol, 33(9), 975–977, doi: 10.1200/JCO.2014.58.2007.

Simon, R. (2018). **New designs for basket clinical trials in oncology**. Journal of Biopharmaceutical Statistics, 28(2), 245–255, doi: 10.1080/10543406.2017.1372779.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). **Bayesian approaches to clinical trials and health-care evaluation**. John Wiley & Sons, Chichester.

Subbiah, V., Kreitman, R. J., Wainberg, Z. A., Gazzah, A., Lassen, U., Stein, A., Wen, P. Y., Dietrich, S., de Jonge, M. J., Blay, J.-Y., et al. (2023). **Dabrafenib plus trametinib in BRAFV600E-mutated rare cancers: the phase 2 ROAR trial**. Nature Medicine, 29(5), 1103–1112, doi: 10.1038/s41591-023-02321-8.

Subbiah, V., Lassen, U., Élez, E., Italiano, A., Curigliano, G., Javle, M., de Braud, F., Prager, G. W., Greil, R., Stein, A., et al. (2020). **Dabrafenib plus trametinib in patients with BRAFV600E-mutated biliary tract cancer (ROAR): A phase 2, open-label, single-arm, multicentre basket trial**. Lancet Oncol, 21(9), 1234–1243, doi: 10.1016/S1470-2045(20)30321-1.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). **Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes**. Statistics in Medicine, 22(5), 763–780, doi: 10.1002/sim.1399.

Thompson, L., Chu, J., Xu, J., Li, X., Nair, R., and Tiwari, R. (2021). **Dynamic borrowing from a single prior data source using the conditional power prior**. Journal of Biopharmaceutical Statistics, 31(4), 403–424, doi: 10.1080/10543406.2021.1895190.

Wickham, H. (2011). **testthat: Get started with testing**. The R Journal, 3(1), 5–10, doi: 0.32614/RJ-2011-002.

Wickham, H. (2023). **testthat: Unit testing for r**. R package version 3.2.0, `https://CRAN.R-project.org/package=testthat`.

Wojciekowski, S. (2022). **bhmbasket: Bayesian hierarchical models for basket trials**. R package version 0.9.5, `https://CRAN.R-project.org/package=bhmbasket`.

Zhou, H., Liu, F., Wu, C., Rubin, E. H., Giranda, V. L., and Chen, C. (2019). **Optimal two-stage designs for exploratory basket trials**. Contemporary Clinical Trials, 85, 105807, doi: 10.1016/j.cct.2019.06.021.

Zhou, H., Shen, R., Wu, S., and He, P. (2023). **A bayesian basket trial design using local power prior**. arXiv, doi: 10.48550/arXiv.2312.15352.

Chapter 8

Personal Contribution and Publications

Parts of this dissertation are covered in the following publication:

Baumann, L., Krisam, J., and Kieser, M. (2022). **Monotonicity conditions for avoiding counterintuitive decisions in basket trials**. Biometrical Journal, 64(5), 934-947, doi: 10.1002/bimj.202100287.

Parts of the results of Section 3.3.3 and the discussion in Section 4.2 are covered in this manuscript. Note that the setup of the systematic investigation and the examples in the manuscript were adapted to align more closely with the parameters of the simulation study in Fujikawa et al. (2020). My contribution to this publication was: development of the methods, writing the R code for the systematic investigation and drafting the manuscript.

Parts of this dissertation are covered in the following submitted manuscripts which are available as preprints on arXiv:

Baumann, L., Sauer, L., and Kieser, M. (2024). **A basket trial design based on power priors**. Submitted to Statistics in Biopharmaceutical Research, doi: 10.48550/arXiv.2309.06988.

Parts of the results of Section 3.1 and 3.2 as well as the discussion in Section 4.1 are covered in this manuscript. My contribution to this publication was: development of most methods, writing the R code for the comparison study and drafting the manuscript. Lukas Sauer derived Equation (3.3).

Baumann, L. (2024). **baskexact: An R package for analytical calculation of basket trial operating characteristics**. Submitted to SoftwareX, doi: 10.48550/arXiv.2403.17510

This manuscript contains a description of the R package baskexact and thus covers parts of the results in Section 3.4. I wrote the R package and the manuscript.

The following R packages related to the thesis were published on CRAN:

Baumann, L. (2024). **baskexact: Analytical calculation of basket trial operating characteristics**, R package version 1.0.1, `https://CRAN.R-project.org/package=baskexact`.

Baumann, L. (2024). **basksim: Simulation-based calculation of basket trial operating characteristics**, R package version 1.0.0, `https://CRAN.R-project.org/package=basksim`.

Other personal publications, not related to the thesis:

Asenbaum, U., Nolz, R., Puchner, S.B., Schoster, T., Baumann, L., Furtner J., Zimpfer D., Laufer, G., Loewe C., and Sandner, S.E. (2020). **Coronary artery bypass grafting and perioperative stroke: imaging of atherosclerotic plaques in the ascending aorta with ungated high-pitch CT-angiography**. Sci Rep, 10, 13909.

Baumann, L., Pilz, M., and Kieser, M. (2022). **blindrecalc - An R package for blinded sample size recalculation**. The R Journal, 14(1), 137-145.

Baumgartner, J., Litvan, Z., Koch, M., Hinterbuchinger, B., Friedrich, F., Baumann, L., and Mossaheb, N. (2020). **Metacognitive beliefs in individuals at risk for psychosis: a systematic review and meta-analysis of sex differences**. Neuropsychiatry, 34(3), 108-115.

Braunschmid, T., Hartig, N., Baumann, L., Dauser, B., and Herbst, F. (2017). **Influence of multiple stapler firings used for rectal division on colorectal anastomotic leak rate**. Surg Endosc, 31(12), 5318-5326.

El Barbari, J. S., Schnetzke, M., Bergmann, M. B., Baumann, L., Vetter, S. Y., Swartman, B., Grützner, P. A., and Franke, J. (2023). **Vascular impulse technology versus elevation for reducing the swelling of upper and lower extremity joint fractures**. Sci Rep, 13, 661.

Denkinger, C. M., Janssen, M., Schaekel, U., Gall, J., Leo, A., Stelmach, P., Weber, S. F., Krisam, J., Baumann, L., Stermann, J., Merle, U., Weigand, M. A., Nusshag, C., Bullinger, L., Schrenzenmeier, J. F., Bornhäuser, M., Alakel, N., Witzke, O., Wolf, T., Vehreschild, M. J. G. T., Schmiedel, S., Addo, M. M., Herth, F., Kreuter, M., Tepasse, P. R., Hertenstein, B., Hänel M., Morgner, A., Kiehl, M., Hopfer, O., Wattad, M., A., Schimanski, C. C., Celik, C., Pohle, T., Ruhe, M., Kern, W. V., Schmitt, A., Lorenz, H. M., Souto-Carneiro, M., Gaeddert, M., Halama, N., Meuer, S., Kräusslich, H. G., Müller, B., Schnitzler, P., Parthé, S., Bartenschläger, R., Gronkowski, M., Klemmer, J., Schmitt M., Dreger, P., Kriegsmann, K., Schlenk, R. F., and Mueller-Tidow, C. (2023). **Anti-SARS-CoV-2 antibody-containing plasma improves outcome in patients with hematologic or solid cancer and severe COVID-19: a randomized clinical trial**. Nature Cancer 4(1), 96-107.

Dörr, C., Kietaibl, C., Dörr, K., Hagmann, M., Baumann, L., Kimberger, O., Ullrich, R., Markstaller, K., and Klein, K.U. (2018). **Impact of CPAP on forehead near-infrared spectroscopy measurements in patients with acute respiratory failure: truth or illusion**. J Neurosurg Anesthesiol, 31(4), 406-412.

Eskandary, F., Regele, H., Baumann, L., Bond, G. , Kozakowski, N., Warhmann, M., Hidalgo, L. G., Haslacher, H., Kaltenecker, C. C., Aretin, M. B., Oberbauer, R., Posch, M., Staudenherz, A., Reeve, J., Halloran, P. F., and Böhmig G.A. (2018). **A randomized trial of bortezomib in late antibody-mediated rejection (BORTEJECT)**. Clin J Am Soc Nephrol, 29(2), 591-605.

Farr, A. Stolz, M., Baumann, L., Bago-Horvath, Z., Oppolzer, E., Pfeiler G., Seifert, M., and Singer C.F. (2017). **The effect of obesity on pathological complete response and survival in breast cancer patients receiving uncapped doses of neoadjuvant anthracycline-taxane-based chemotherapy**. The Breast, 33, 153-158.

Hoffmann, P., Krueger, J., Bashlekova, T., Rupp, C., Baumann, L., and Gauss, A. (2021). **Pregnancy with inflammatory bowel disease: Outcomes for mothers and their children at a European tertiary care center**. J Obstet Gynaecol Res, 48(3), 621-633.

Jaramillo, S., Krisam, J., Le Cornet, L., Kratzmann, M., Baumann, L., Sauer, T., Crysandt, M., Rank, A., Behringer, D., Teichmann, L., Görner, M., Trappe, R.U., Röllig, C., Krause, S., Hanoun, M., Hopfer, O., Held, G., Buske, S., Fransecky, L., Kayser, S., Schliemann, C., Schaefer-Eckart, K., Al-Fareh, Y., Schubert, J., Geer, T., Kaufmann, M., Brecht, A., Niemann, D., Kieser, M., Bornhäuser, M., Platzbecker, U., Serve, H., Baldus, C. D., Müller-Tidow, C., and Schlenk, R.F. (2021). **Rationale and design of the 2 by 2 factorial design GnG-trial: a randomized phase-III study to compare two schedules of gemtuzumab ozogamicin as adjunct to intensive induction therapy and to compare double-blinded intensive postremission therapy with or without glasdegib in older patients with newly diagnosed AML**. Trials, 22, 765.

Kapsner, L. A., Balbach, E. L., Laun, F. B., Baumann, L., Ohlmeyer S., Uder, M., Bickelhaupt, S., and Enkel E. (2023). **Prevalence and influencing factors for artifact development in breast MRI-derived maximum intensity projections**. Acta Radiologica, 64(11), 2881-2890.

Keilani, M., Hasenöhrl, T., Baumann, L., Ristl, R., Schwarz, M., Marhold, M., Sedghi Komandj, T., and Crevenna R. (2017). **Effects of resistance exercise in prostate cancer patients: a meta-analysis**. Support Care Cancer, 25(9), 2953-2968.

Mack, C. E, Klaiber, U., Sauer, P., Kohlhas, L., Baumann, L., Martin, E., Mehrabi, A., Buchler, M. W., and Hackert, T. (2023). **Protocol of a randomised controlled phase II clinical trial investigating preoperative endoscopic injection of botulinum**

toxin into the sphincter of oddi to reduce bile leakage after hepatic resection: the PREBOT-II trial. BMJ Open, 13(9), e065727.

Mylonas, G., Schranz, M., Georgopoulos, M., Sacu, S., Deak, G., Reumueller, A., Baumann, L., and Schmidt-Erfurth, U. (2020). **Are there funduscopic and optical coherence tomography preoperative characteristics to predict surgical difficulty of epiretinal membrane removal?**. Curr Eye Res, 45(8), 1012-1016.

Mylonas, G., Schranz, M., Scholda, C., Karst, S., Reiter, G., Baumann, L., Schmidt-Erfurth, U., and Kriechbaum, K. (2020). **Response of retinal sensitivity to intravitreal anti-angiogenic bevacizumab and triamcinolone acetonide for patients with diabetic macular edema over one year**. Curr Eye Res, 45(9), 1107-1113.

Nemecek, R., Huber, P., Schur, S., Masel, E.K., Baumann, L., Hoeller, C., Watzke, H., and Binder, M. (2019). **Telemedically augmented palliative care : Empowerment for patients with advanced cancer and their family caregivers**. Wien Klin Wochenschr, 131, 620-626.

Patry, C., Cordts, S., Baumann, L., Höcker, B., Fichtner, A., Ries, M., and Tönshoff, B. (2022). **Publication rate and research topics of studies in pediatric kidney transplantation**. Pediatr Transplant, 26(4), e14262.

Patry, C., Höcker, B., Dello Strologo, L., Baumann, L., Grenda, R., Peruzzi, L., Oh, J., Pape, L., Weber, L. T., Weitz, M., Awan, A., Carraro, A., Zirngibl, M., Hansen, M., Müller, D., Bald, M., Pecqueuex, C., Krupka, K., Fichtner, A., Tönshoff, B., and Nyarangi-Dix, J. (2022). **Timing of reconstruction of the lower urinary tract in pediatric kidney transplant recipients: A CERTAIN multicenter analysis of current practice**. Pediatr Transplant, 26(6), e14328.

Pfisterer, N., Dexheimer, C., Fuchs, E.M., Bucsics, T., Schwabl, P., Mandorfer, M., Gessl, I., Sandrieser, L., Baumann, L., Riedl, F., Scheiner, B., Pachofszky, T., Dolak, W., Schrutka-Kölbl, C., Ferlitsch, A., Schöniger-Hekele, M., Peck, M., Trauner, M., Madl, C., and Reiberger, T. (2018). **Beta-blockers do not increase efficacy of band ligation**

in primary prophylaxis but improve survival in secondary prophylaxis of variceal bleeding. Aliment Pharmacol Ther, 47(7), 966-979.

Pohl, M., Baumann, L., Behnisch, R., Kirchner, M., Krisam, J., and Sander, A. (2021). **Estimands-a basic element for clinical trials. Part 29 of a series on evaluation of scientific publications**. Dtsch Arztebl Int, 118(51-52), 883–888.

Prager, G., Unseld, M., Waneck, F., Mader, R., Wrba, F., Raderer, M., Füreder, T., Staber, P., Jäger, U., Kieler, M., Bianconi, D., Hoda, M.A., Baumann, L., Reithaller, A., Berger, W., Grimm, C., Kölbl, H., Sibilia, M., Müllauer, L., and Zielinski, C. (2019). **Results of the extended analysis for cancer treatment (EXACT) trial: a prospective translational study evaluating individualized treatment regimens in oncology**. Oncotarget, 10(9), 942-952.

Primas, C., Hopf, G., Reinisch, S., Baumann, L., Novacek, G., Reinisch, W., and Vogelsang, H. (2021). **Role of fecal calprotection in predicting endoscopic recurrence in postoperative Crohn's disease**. Scand J Gastroenterol, 56(10), 1169-1174.

Reiter, G., Told, R., Baumann, L., Sacu, S., Schmidt-Erfurth, U., and Pollreisz, A. (2020). **Investigating a growth prediction model in advanced age-related macular degeneration with solitary geographic atrophy using quantitative autofluorescence**. Retina, 40(9), 1657-1664.

Reiter, G. S., Told, R., Schranz, M., Baumann, L., Mylonas, G., Sacu, S., Pollreisz, A., and Schmidt-Erfurth, U. (2020). **Subretinal drusenoid deposits and photoreceptor loss detecting global and local progression of geographic atrophy by SD-OCT imaging**. Invest Ophthalmol Vis Sci, 61(6), 11.

Reiter, G., Told, R., Schlanitz, F., Baumann, L., Schmidt-Erfurth, U., and Sacu, S. (2019). **Longitudinal association between drusen volume and retinal capillary perfusion in intermediate age-related macular degeneration**. Invest Ophthalmol Vis Sci, 60(7), 2503-2508.

Reiter, G., Told, R., Schlanitz, F., Bogunovic, H., Baumann, L., Sacu, S., Schmidt-Erfurth, U., and Pollreisz, A. (2019). **Impact of drusen volume on quantitative fundus**

autofluorescence in early and intermediate age-related macular degeneration over time. Invest Ophthalmol Vis Sci, 60(6), 1937-1942.

Ristau, J., Hörner-Rieber, J., Buchele, C., Klüter, S., Jäkel, C., Baumann, L., Andratschke, N., Garcia Schüler, H., Guckenberger, M., Li, M., Niyazi, M., Belka, C., Herfarth, K., Debus, J., and Koerber, S. A. (2022). **Stereotactic MRI-guided radiation therapy for localized prostate cancer (SMILE): a prospective, multicentric phase-II-trial**. Radiation Oncology, 17(1), 1-8.

Schlanitz, F., Baumann, B., Sacu, S., Baumann, L., Pircher, M., Hitzenberger, C.K., and Schmidt-Erfurth, U.M. (2019). **Impact of drusen and drusenoid retinal pigment epithelium elevation size and structure on the integrity of the retinal pigment epithelium layer**. Br J Ophthalmol, 103(2), 227-232.

Schuhn, A., Tobar, T. W., Gahlawat, A. W., Hauke, J., Baumann, L., Okun, J. G., and Nees, J. (2022). **Potential of blood-based biomarker approaches in endometrium and breast cancer: a case-control comparison study**. Arch Gynecol Obstet, 306(5), 1623-1632.

Silberhumer, G. R., Györi, G., Brugger, J., Baumann, L., Zehetmayer, S., Soliman, T., and Berlakovich, G. (2023). **MELD-Na alterations on the liver transplant waiting list and their impact on listing outcome**. J Clin Med, 12(11), 3763.

Told, R., Georgopoulos M., Reiter, G.S., Wassermann, L., Aliyeva, L., Baumann, L., Abela-Formanek, C., Pollreisz, A., Schmidt-Erfurth, U., and Sacu, S. (2020). **Intraretinal microvascular changes after ERM and ILM peeling using SSOCTA**. PloS One, 15(12), e0242667.

Unseld, M., Krammer, K., Lubowitzki, S., Jachs, M., Baumann, L., Vyssoki, B., Riedel, J., Puhr, H., Zehentgruber, S., Prager, G., Masel, E.K., Preusser, M., Jaeger, U., and Gaiger, A. (2019). **Screening for post-traumatic stress disorders in 1017 cancer patients and correlation with anxiety, depression, and distress**. Psychooncology, 28(12), 2382-2388.

Unseld, M., Mader, R., Baumann, L., Veraar, C., Wrba, F., Waneck, F., Kieler, M., Bianconi, D., Berger, W., Sibilia, M., Müllauer, M., Zielinski, C., and Prager, G. W. (2018). **Feasibility of personalized treatment concepts in gastrointestinal malignancies: Sub-group results of prospective clinical phase II trial EXACT**. Chin J Cancer Res, 30(5), 508-515.

Wick, A., Sander, A., Koch, M., Bendszus, M., Combs, S., Haut, T., Dormann, A., Walter, S., Pertz, M., Merkle-Lock, J., Limprecht, R., Baumann, L., Kieser, M., Sahm, F., Schlegel, U., Winkler, F., Platten, M., Wick, W., and Kessler, T. (2022). **Improvement of functional outcome for patients with newly diagnosed grade 2 or 3 gliomas with co-deletion of 1p/19q–IMPROVE CODEL: the NOA-18 trial**. BMC Cancer, 22, 645.

Appendix

R-Code

With the R code below, the results of this thesis can be reproduced using the two R packages baskexact and basksim which were developed as part of this thesis and are described in Section 3.4 and Section 3.5. The source code of the two R packages can be viewed on GitHub (`https://github.com/lbau7/baskexact` and `https://github.com/lbau7/basksim`). Both R packages are also available on CRAN.

```
# Load packages
library(baskexact)
library(basksim)
# devtools::install_github("https://github.com/lbau7/bhmbasket")
library(bhmbasket)
library(doFuture)
library(extraDistr)
library(extrafont)
library(ggplot2)
library(gridExtra)
library(latex2exp)
library(progressr)
library(tidyverse)
library(viridis)

### Figure 1
# Prior and posterior
x <- seq(0, 1, by = 0.001)
yprior <- dbeta(x, 1, 1)
yposterior <- dbeta(x, 1 + 5, 1 + 15)

data <- data.frame(x = rep(x, 2), y = c(yprior, yposterior),
  type = factor(c(rep("prior", 1001), rep("posterior", 1001)),
    levels = c("prior", "posterior")))
```

```r

p1 <- ggplot(data) +
  geom_line(aes(x = x, y = y, colour = type)) + theme_bw() +
  theme(legend.title = element_blank(),
    legend.position = c(0.8, 0.85)) +
  ylab("f(x)")

# Posterior Predictive
x <- 0:10
ypostpred <- dbbinom(x, size = 10, alpha = 6, beta = 16)

data <- data.frame(x, ypostpred)
p2 <- ggplot(data, aes(x = x, y = ypostpred)) +
  geom_bar(stat = "identity", fill = "deepskyblue3") +
  scale_x_continuous(breaks = 0:10) + theme_bw() +
  xlab(TeX("$\\tilde{r}$")) + ylab(TeX("f(\\tilde{r})$"))

grid.arrange(p1, p2, ncol = 2)

### Figure 2
plot_weights(
  design = setupOneStageBasket(k = 3, p0 = 0.2),
  n = 20, r1 = 10,
  weight_fun = weights_fujikawa,
  weight_params = list(epsilon = c(0.5, 1:4), tau = 0)
) + scale_colour_discrete(name = TeX("$\\epsilon"))

### Figure 3
plot_old <- plot_weights(
  design = setupOneStageBasket(k = 3, p0 = 0.2),
  n = 20, r1 = 10,
  weight_fun = weights_fujikawa,
  weight_params = list(epsilon = c(1.25, 2),
    logbase = c(exp(1), 2), tau = 0))

plotdata <- as_tibble(plot_old$data)
plotdata_new <- plotdata %>% filter(
  (param1 == "2" & param2 == "2.71828182845905") |
    (param1 == "1.25" & param2 == "2")
)

ggplot(plotdata_new, aes(x = r, y = weight)) +
  geom_line(aes(col = param2)) +
  theme_bw() +
  scale_color_discrete(
    labels = unname(c(TeX("$log_2, \\epsilon = 1.25$"),
      TeX("$log, \\epsilon = 2$"))),
    name = "")

### Figure 4
plot_weights(
  design = setupOneStageBasket(k = 3, p0 = 0.2),
  n = 20, r1 = 10,
```

```
  weight_fun = weights_cpp,
  weight_params = list(a = 1:3, b = 1:3)
)

### Figure 5
plot_a <- plot_weights(
  design = setupOneStageBasket(k = 3, p0 = 0.2),
  n = 20, r1 = 10,
  weight_fun = weights_mml
)

plot_b <- plot_weights(
  design = setupOneStageBasket(k = 3, p0 = 0.2),
  n = 20, r1 = 10,
  weight_fun = weights_cpp,
  weight_params = list(a = 8, b = 8.5)
)

plotdata <- data.frame(
  rbind(
    cbind("Method" = "MML", plot_a$data),
    cbind("Method" = "CPP (a = 8, b = 8.5)", plot_b$data)
  ))

ggplot(plotdata, aes(x = r, y = weight)) +
  geom_line(aes(col = Method)) +
  theme_bw()

### Figure 6
hlab <- "h (\u03B5* = 2.5)"
jsdlab <- "JSD (\u03B5* = 1)"
r1lab <- "r\u2081"
r2lab <- "r\u2082"
windowsFonts(font = windowsFont("Microsoft Sans Serif"))

r1 <- r2 <- 0:20
h <- Vectorize(function(x, y) basksim:::diff_all(n = 20,
  r = c(10, x, y), epsilon = 2.5))
zh <- outer(r1, r2, h)

jsd <- Vectorize(function(x, y) {
  shape1 <- c(1 + 10, 1 + x, 1 + y)
  shape2 <- c(1 + 20 - 10, 1 + 20 - x, 1 + 20 - y)
  shape <- rbind(shape1, shape2)

  basksim:::jsd_global(shape, epsilon = 1)
})
zjsd <- outer(r1, r2, jsd)

par(mfrow = c(1, 2), mai = c(0.3, 0.3, 0.3, 0.3), family = "font")
persp(
  r1, r2, zh,
  theta = 30, phi = 20, col = "cyan", ticktype = "detailed",
```

```
  main = hlab, zlab = "weight", xlab = r1lab, ylab = r2lab,
  cex.lab = 1, cex.axis = 0.7
)
persp(
  r1, r2, zjsd,
  theta = 30, phi = 20, col = "cyan", ticktype = "detailed",
  main = jsdlab, zlab = "weight", xlab = r1lab, ylab = r2lab,
  cex.lab = 1, cex.axis = 0.7
)

### Comparison Study
handlers(global = TRUE) # show progressbar
plan(multisession, workers = 10)

# Define Scenarios
scenarios <- data.frame(
  "Global_Null" = c(0.15, 0.15, 0.15, 0.15),
  "Global_Alternative" = c(0.4, 0.4, 0.4, 0.4),
  "One_in_the_Middle" = c(0.4, 0.4, 0.3, 0.5),
  "Linear" = c(0.15, 0.25, 0.35, 0.45),
  "Good_Nugget" = c(0.15, 0.15, 0.15, 0.4),
  "Bad_Nugget" = c(0.15, 0.4, 0.4, 0.4),
  "Half" = c(0.15, 0.15, 0.4, 0.4)
)

# Calculations with baskexact
design_exact <- setupOneStageBasket(k = 4, p0 = 0.15,
  shape1 = 1, shape2 = 1)

# Parameter and argument list for all methods
params_exact <- list(
  "cpp" = list(
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(from = 0.5, to = 3, by = 0.5),
      b = seq(from = 0.5, to = 3, by = 0.5)
    ),
    globalweight_fun = NULL,
    globalweight_params = list()
  ),
  "cppglobal" = list(
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(from = 0.5, to = 3, by = 0.5),
      b = seq(from = 0.5, to = 3, by = 0.5)
    ),
    globalweight_fun = globalweights_diff,
    globalweight_params = list(eps_global =
      seq(from = 0.5, to = 3, by = 0.5))
  ),
  "cppnex" = list(
    weight_fun = weights_cpp,
    weight_params = list(
```

```
      a = seq(from = 0.5, to = 3, by = 0.5),
      b = seq(from = 0.5, to = 3, by = 0.5)
    ),
    globalweight_fun = globalweights_fix,
    globalweight_params = list(w =
      seq(from = 0.1, to = 0.9, by = 0.1))
  ),
  "fujikawa" = list(
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = seq(from = 0.5, to = 3, by = 0.5),
      tau = seq(from = 0, to = 0.5, by = 0.1)
    ),
    globalweight_fun = NULL,
    globalweight_params = list()
  ),
  "mml" = list(
    weight_fun = weights_mml,
    weight_params = list(),
    globalweight_fun = NULL,
    globalweight_params = list()
  )
)

# Run optimisation
res_exact <- list()
methods_exact <- c("cpp", "cppglobal", "cppnex", "fujikawa", "mml")
for (i in methods_exact) {
  res_exact[[i]] <- baskexact::opt_design(
    design = design_exact,
    n = 20, alpha = 0.05,
    weight_fun = params_exact[[i]]$weight_fun,
    weight_params = params_exact[[i]]$weight_params,
    globalweight_fun = params_exact[[i]]$globalweight_fun,
    globalweight_params = params_exact[[i]]$globalweight_params,
    scenarios = scenarios,
    prec_digits = 3
  )
}

# Calculations with basksim - methods not requiring MCMC sampling
# Create simulated data set
scenario_list <- as.list(scenarios)
set.seed(123)
data_list_mat <- lapply(scenario_list,
  function(x) get_data(k = 4, n = 20, p = x, iter = 10000))

# Create design objects
design_bma <- setup_bma(k = 4, p0 = 0.15, shape1 = 1, shape2 = 1)
design_jsdglobal <- setup_jsdglobal(k = 4, p0 = 0.15,
  shape1 = 1, shape2 = 1)
design_mmlglobal <- setup_mmlglobal(k = 4, p0 = 0.15,
  shape1 = 1, shape2 = 1)
```

```
237
238 # Parameter and argument list for all methods
239 params_sim = list(
240   "bma" = list(
241     design = design_bma,
242     design_params = list(pmp0 = seq(from = -4, to = 4, by = 0.5))
243   ),
244   "jsdglobal" = list(
245     design = design_jsdglobal,
246     design_params = list(
247       eps_pair = seq(from = 0.5, to = 3, by = 0.5),
248       eps_all = seq(from = 0.5, to = 3, by = 0.5),
249       tau = seq(from = 0, to = 0.5, by = 0.1),
250       logbase = 2
251     )
252   ),
253   "mmlglobal" = list(
254     design = design_mmlglobal,
255     design_params = list()
256   )
257 )
258
259 # Run optimisation
260 res_sim <- list()
261 methods_sim <- c("bma", "jsdglobal", "mmlglobal")
262 for (i in methods_sim) {
263   res_sim[[i]] <- basksim::opt_design(
264     design = params_sim[[i]]$design,
265     n = 20,
266     alpha = 0.05,
267     design_params = params_sim[[i]]$design_params,
268     scenarios = scenarios,
269     prec_digits = 3,
270     iter = 10000,
271     data = data_list_mat
272   )
273 }
274
275 # Calculations with basksim - methods requiring MCMC sampling
276 # set same seed again to have the same simulated data sets for
277 # all methods
278 set.seed(123)
279 data_list_bhm <- lapply(scenario_list,
280   function(x) get_data(k = 4, n = 20, p = x, iter = 10000,
281     type = "bhmbasket"))
282
283 # Create design objects
284 design_bhm <- setup_bhm(k = 4, p0 = 0.15, p_target = 0.4)
285 design_exnex <- setup_exnex(k = 4, p0 = 0.15)
286
287 # Parameter and argument list for all methods
288 params_mcmc = list(
289   "bhm" = list(
```

```
    design = design_bhm,
    design_params = list(tau_scale =
      seq(from = 0.125, to = 2, length.out = 8)),
    workers = 10
  ),
  "exnex" = list(
    design = design_exnex,
    design_params = list(
      tau_scale = seq(from = 0.125, to = 2, length.out = 8),
      w = seq(from = 0.1, to = 0.9, by = 0.1)
    ),
    workers = 11
  )
)

# Run optimisation
res_mcmc <- list()
# Optimisation of EXNEX may take several days
methods_mcmc <- c("bhm", "exnex")
for (i in methods_mcmc) {
  # calling registerdoFuture is necessary for parallelisation in
  # bhmbasket, since an older version of the doFuture package is
  # used there
  # the number of cores cannot be changed for exact reproducibility of
  # the results due to the way bhmbasket chunks tasks when it is run
  # in parallel
  registerDoFuture()
  plan(multisession, workers = params_mcmc[[i]]$workers)
  # setting the seed is necessary since MCMC sampling
  # introduces randomness
  set.seed(234)

  res_mcmc[[i]] <- basksim::opt_design(
    design = params_mcmc[[i]]$design,
    n = 20, alpha = 0.05,
    design_params = params_mcmc[[i]]$design_params,
    scenarios = scenarios,
    prec_digits = 3, iter = 10000,
    data = data_list_bhm
  )
}

# Combine all results
res <- c(res_exact, res_sim, res_mcmc)

# Best results and tuning parameter values per method
# Table 3 & Table 4
lapply(res, function(x) x[1, ])

# Rejection rates and mean posterior means - baskexact
params_exact_opt <- list(
  "cpp" = list(
    lambda = res_exact$cpp$Lambda[1],
```

```
    weight_fun = weights_cpp,
    weight_params = list(
      a = res_exact$cpp$a[1],
      b = res_exact$cpp$b[1]
    ),
    globalweight_fun = NULL,
    globalweight_params = list()
  ),
  "cppglobal" = list(
    lambda = res_exact$cppglobal$Lambda[1],
    weight_fun = weights_cpp,
    weight_params = list(
      a = res_exact$cppglobal$a[1],
      b = res_exact$cppglobal$b[1]
    ),
    globalweight_fun = globalweights_diff,
    globalweight_params = res_exact$cppglobal$eps_global[1]
  ),
  "cppnex" = list(
    lambda = res_exact$cppnex$Lambda[1],
    weight_fun = weights_cpp,
    weight_params = list(
      a = res_exact$cppnex$a[1],
      b = res_exact$cppnex$b[1]
    ),
    globalweight_fun = globalweights_fix,
    globalweight_params = res_exact$cppnex$w[1]
  ),
  "fujikawa" = list(
    lambda = res_exact$fujikawa$Lambda[1],
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = res_exact$fujikawa$epsilon[1],
      tau = res_exact$fujikawa$tau[1]
    ),
    globalweight_fun = NULL,
    globalweight_params = list()
  ),
  "mml" = list(
    lambda = res_exact$mml$Lambda[1],
    weight_fun = weights_mml,
    weight_params = list(),
    globalweight_fun = NULL,
    globalweight_params = list()
  )
)

scenarios_toer <- scenario_list[c(1, 4, 5, 6, 7)]
scenarios_pow <- scenario_list[c(2, 3)]

toer_exact <- list()
for (i in methods_exact) {
  toer_exact[[i]] <- t(sapply(scenarios_toer, function(x)
```

```
    unlist(unname(baskexact::toer(
      design = design_exact,
      p1 = x, n = 20,
      lambda = params_exact_opt[[i]]$lambda,
      weight_fun = params_exact_opt[[i]]$weight_fun,
      weight_params = params_exact_opt[[i]]$weight_params,
      globalweight_fun = params_exact_opt[[i]]$globalweight_fun,
      globalweight_params = params_exact_opt[[i]]$globalweight_params,
      results = "group"
    )))))
}

pow_exact <- list()
for (i in methods_exact) {
  pow_exact[[i]] <- t(sapply(scenarios_pow, function(x)
    unlist(unname(baskexact::pow(
      design = design_exact,
      p1 = x, n = 20,
      lambda = params_exact_opt[[i]]$lambda,
      weight_fun = params_exact_opt[[i]]$weight_fun,
      weight_params = params_exact_opt[[i]]$weight_params,
      globalweight_fun = params_exact_opt[[i]]$globalweight_fun,
      globalweight_params = params_exact_opt[[i]]$globalweight_params,
      results = "group"
    )))))
}

# Combind toer and pow
rej_exact <- lapply(1:5, function(x)
  rbind(toer_exact[[x]], pow_exact[[x]]))
names(rej_exact) <- names(toer_exact)

estim_exact <- list()
for (i in methods_exact) {
  estim_exact[[i]] <- t(sapply(scenario_list, function(x)
    unlist(unname(baskexact::estim(
      design = design_exact,
      p1 = x, n = 20,
      lambda = params_exact_opt[[i]]$lambda,
      weight_fun = params_exact_opt[[i]]$weight_fun,
      weight_params = params_exact_opt[[i]]$weight_params,
      globalweight_fun = params_exact_opt[[i]]$globalweight_fun,
      globalweight_params = params_exact_opt[[i]]$globalweight_params,
    )))))
}
# Keep estimates, remove MSEs
estim_exact <- lapply(estim_exact, function(x) x[, 1:4])

# Rejection rates and mean posterior means - basksim without MCMC
details_bma <- t(sapply(1:7, function(x)
  unlist(unname(basksim::get_details(
    design = design_bma,
    n = 20, p1 = scenario_list[[x]],
```

```
    lambda = res_sim$bma$Lambda[1],
    pmp0 = res_sim$bma$pmp0[1],
    iter = 10000, data = data_list_mat[[x]]
  )))))

details_jsdglobal <- t(sapply(1:7, function(x)
  unlist(unname(basksim::get_details(
    design = design_jsdglobal,
    n = 20, p1 = scenario_list[[x]],
    lambda = res_sim$jsdglobal$Lambda[1],
    eps_pair = res_sim$jsdglobal$eps_pair[1],
    tau = res_sim$jsdglobal$tau[1],
    eps_all = res_sim$jsdglobal$eps_all[1],
    logbase = 2,
    iter = 10000, data = data_list_mat[[x]]
  )))))

details_mmlglobal <- t(sapply(1:7, function(x)
  unlist(unname(basksim::get_details(
    design = design_mmlglobal,
    n = 20, p1 = scenario_list[[x]],
    lambda = res_sim$mmlglobal[1],
    iter = 10000, data = data_list_mat[[x]]
  )))))

# Rejection rates and mean posterior means - basksim with MCMC
set.seed(345)
registerDoFuture()
plan(multisession, workers = 10)
details_bhm <- t(sapply(1:7, function(x)
  unlist(unname(basksim::get_details(
    design = design_bhm,
    n = 20, p1 = scenario_list[[x]],
    lambda = res_mcmc$bhm$Lambda[1],
    tau_scale = res_mcmc$bhm$tau_scale[1],
    iter = 10000, data = data_list_bhm[[x]]
  )))))

set.seed(345)
registerDoFuture()
plan(multisession, workers = 10)
details_exnex <- t(sapply(1:7, function(x)
  unlist(unname(basksim::get_details(
    design = design_exnex,
    n = 20, p1 = scenario_list[[x]],
    lambda = res_mcmc$exnex$Lambda[1],
    tau_scale = res_mcmc$exnex$tau_scale[1],
    w = res_mcmc$exnex$w[1],
    iter = 10000, data = data_list_bhm[[x]]
  )))))

# All rejection rates
rej_all <- c(rej_exact,
```

```
  list("bhm" = details_bhm[, 1:5]),
  list("exnex" = details_exnex[, 1:5]),
  list("jsdglobal" = details_jsdglobal[, 1:5]),
  list("bma" = details_bma[, 1:5]),
  list("mmlglobal" = details_mmlglobal[, 1:5])
)

# All posterior means
estim_all <- c(estim_exact,
  list("bhm" = details_bhm[, 6:9]),
  list("exnex" = details_exnex[, 6:9]),
  list("jsdglobal" = details_jsdglobal[, 6:9]),
  list("bma" = details_bma[, 6:9]),
  list("mmlglobal" = details_mmlglobal[, 6:9])
)

# Mean posterior mean plot - Figure 7
estim_df <- do.call(rbind, estim_all)
rownames(estim_df) <- NULL
colnames(estim_df) <- paste("Basket", 1:4)
scenario_names <- c("Global Null", "Global Alt", "One in the Middle",
  "Linear", "Good Nugget", "Bad Nugget", "Half")
estim_df <- data.frame(
  "Method" = rep(names(estim_all), each = 7),
  "Scenario" = rep(scenario_names, times = length(estim_all)),
  estim_df
)

resprates <- data.frame(scenario_names, t(scenarios))
colnames(resprates) <- c("Scenario", paste("Basket", 1:4))
resprates <- as_tibble(resprates) %>%
  pivot_longer(-1, names_to = "Basket")

estim_long <- estim_df %>% pivot_longer(-(1:2), names_to = "Basket",
  values_to = "Mean") %>%
  mutate(
    Basket = case_when(
      Basket == "Basket.1" ~ "Basket 1",
      Basket == "Basket.2" ~ "Basket 2",
      Basket == "Basket.3" ~ "Basket 3",
      Basket == "Basket.4" ~ "Basket 4"
    ),
    Scenario = factor(Scenario, levels = c("Global Null", "Global Alt"
      ,
      "One in the Middle", "Linear", "Good Nugget", "Bad Nugget", "
        Half"))
  ) %>%
  left_join(resprates, by = c("Scenario", "Basket"))

ggplot(estim_long, aes(x = 1, y = Mean, fill = Method)) +
  geom_bar(position = position_dodge2(), stat = "identity") +
  geom_hline(aes(yintercept = value), col = "green", linewidth = 1) +
  scale_fill_viridis(discrete = T) +
```

```
  facet_grid(Scenario ~ Basket) +
  theme_bw() +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
  ) +
  xlab("") +
  ylab("mean posterior mean")


# Sensitivity Analyses - Figure 8
heat_bhm <- ggplot(res$bhm) +
  geom_tile(aes(y = tau_scale, x = 1, fill = Mean_ECD)) +
  scale_y_continuous(
    breaks = seq(0.125, 2, length.out = 4),
    labels = function(x) round(x, 1),
    expand = c(0, 0)) +
  scale_x_continuous(expand = c(0, 0), breaks = NULL) +
  scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
    option = "A") +
  theme_bw() + xlab("") + ylab(TeX("$\\phi")) + ggtitle("BHM") +
  theme(legend.position = "none") + coord_flip()

heat_bma <- ggplot(res$bma) +
  geom_tile(aes(y = pmp0, x = 1, fill = Mean_ECD)) +
  scale_y_continuous(
    breaks = round(seq(-4, 4, by = 1), 3),
    expand = c(0, 0)) +
  scale_x_continuous(expand = c(0, 0), breaks = NULL) +
  scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
    option = "A") +
  theme_bw() + xlab("") + ylab(TeX("$\\psi")) + ggtitle("BMA") +
  theme(legend.position = "none") +
  coord_flip()

heat_cpp <- ggplot(res$cpp) +
  geom_tile(aes(x = a, y = b, fill = Mean_ECD)) +
  scale_x_continuous(breaks = seq(0.5, 3, by = 1),
    expand = c(0, 0)) +
  scale_y_continuous(breaks = seq(0.5, 3, by = 0.5),
    expand = c(0, 0)) +
  scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
    option = "A") +
  theme_bw() + ggtitle("CPP") + theme(legend.position = "none")

heat_cppglobal <- res$cppglobal %>%
  mutate(eps_global = paste(TeX("$\\epsilon^*"),
    eps_global, sep = ": ")) %>%
  ggplot() +
  geom_tile(aes(x = a, y = b, fill = Mean_ECD)) +
  facet_wrap(vars(eps_global), ncol = 2,
    labeller = label_parsed) +
  scale_x_continuous(breaks = seq(0.5, 3, by = 1),
```

```
606     expand = c(0, 0)) +
607   scale_y_continuous(breaks = seq(0.5, 3, by = 0.5),
608     expand = c(0, 0)) +
609   scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
610     option = "A") +
611   theme_bw() + ggtitle("CPP-Global") + theme(legend.position = "none")
612
613 heat_cppnex <- ggplot(res$cppnex) +
614   geom_tile(aes(x = a, y = b, fill = Mean_ECD)) +
615   facet_wrap(vars(w), ncol = 3,
616     labeller = labeller(.rows = label_both)) +
617   scale_x_continuous(breaks = seq(0.5, 3, by = 1),
618     expand = c(0, 0)) +
619   scale_y_continuous(breaks = seq(0.5, 3, by = 0.5),
620     expand = c(0, 0)) +
621   scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
622     option = "A") +
623   theme_bw() + ggtitle("CPP-Nex") + theme(legend.position = "bottom")
624
625 heat_exnex <- ggplot(res$exnex) +
626   geom_tile(aes(x = w, y = tau_scale, fill = Mean_ECD)) +
627   scale_y_continuous(
628     breaks = seq(0.125, 2, length.out = 4),
629     labels = function(x) round(x, 1),
630     expand = c(0, 0)) +
631   scale_x_continuous(breaks = seq(0.1, 0.9, by = 0.2),
632     expand = c(0, 0)) +
633   scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
634     option = "A") +
635   theme_bw() + coord_flip() + ylab(TeX("$\\phi")) + ggtitle("EXNEX") +
636   theme(legend.position = "none")
637
638 heat_fujikawa <- ggplot(res$fujikawa) +
639   geom_tile(aes(x = epsilon, y = tau, fill = Mean_ECD)) +
640   scale_x_continuous(breaks = seq(0.5, 3, by = 1),
641     expand = c(0, 0)) +
642   scale_y_continuous(breaks = seq(0, 0.5, by = 0.1),
643     expand = c(0, 0)) +
644   scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
645     option = "A") +
646   theme_bw() + xlab(TeX("\\epsilon")) + ylab(TeX("\\tau")) +
647   ggtitle("Fujikawa") +
648   theme(legend.position = "none")
649
650 heat_jsdglobal <- res$jsdglobal %>%
651   mutate(eps_all = paste(TeX("$\\epsilon^*"), eps_all,
652     sep = ": ")) %>%
653   ggplot() + geom_tile(aes(x = eps_pair, y = tau, fill = Mean_ECD)) +
654   facet_wrap(vars(eps_all), ncol = 2,
655     labeller = label_parsed) +
656   scale_x_continuous(breaks = seq(0.5, 3, by = 1), expand = c(0, 0)) +
657   scale_y_continuous(breaks = seq(0, 0.5, by = 0.1),
658     expand = c(0, 0)) +
```

```
  scale_fill_viridis_c(name = "Mean ECD", limits = c(3.17, 3.566),
    option = "A") +
  theme_bw() + xlab(TeX("\\epsilon")) + ylab(TeX("\\tau")) +
  ggtitle("JSD-Global") + theme(legend.position = "none")

laymat <- rbind(
  c(1, 1, 2, 2, 3, 3, 4, 4),
  c(5, 5, 2, 2, 3, 3, 4, 4),
  c(6, 6, 6, 6, 7, 7, 7, 7),
  c(6, 6, 6, 6, 7, 7, 7, 7),
  c(6, 6, 6, 6, 7, 7, 7, 7),
  c(6, 6, 6, 6, 7, 7, 7, 7),
  c(6, 6, 6, 6, 7, 7, 7, 7),
  c(NA, 8, 8, 8, 8, 8, 8, NA),
  c(NA, 8, 8, 8, 8, 8, 8, NA),
  c(NA, 8, 8, 8, 8, 8, 8, NA),
  c(NA, 8, 8, 8, 8, 8, 8, NA),
  c(NA, 8, 8, 8, 8, 8, 8, NA),
  c(NA, 8, 8, 8, 8, 8, 8, NA)
)

#heatplot <- arrangeGrob(
heatplot <- arrangeGrob(
  heat_bma, # 1
  heat_exnex, # 2
  heat_fujikawa, # 3
  heat_cpp, # 4
  heat_bhm, # 5
  heat_jsdglobal, # 6
  heat_cppglobal, # 7
  heat_cppnex, # 8
  layout_matrix = laymat
)

# Sensitivity Analyses - Table 6 & 7
res_sens <- list()
for (i in names(res)) {
  temp <- res[[i]]
  temp_mod <- as.data.frame(temp) %>%
    select(-Mean_ECD) %>%
    rowwise() %>%
    mutate(mean_sens = mean(c(Global_Null, Global_Alternative,
    Good_Nugget, Bad_Nugget, Half))) %>%
    arrange(desc(mean_sens))
  res_sens[[i]] <- as.data.frame(temp_mod[1, ])
}

# Sensitivity Analyses - Table 8 & 9
lapply(res, function(x)
  as.data.frame(x)[which.max(as.data.frame(x)$Linear), ])
lapply(res, function(x)
  as.data.frame(x)[which.max(as.data.frame(x)$Bad_Nugget), ])
lapply(res, function(x)
```

```
  as.data.frame(x)[which.max(as.data.frame(x)$Half), ])

### Non-Monotonicity Conditions
## Examples
# Within-Trial Non-Monotonicity
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(5, 5, 5, 6), lambda = 0.99,
  weight_fun = weights_cpp, weight_params = list(a = 1.5, b = 0.5))
# With Pruning
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(5, 5, 5, 6), lambda = 0.99,
  weight_fun = weights_cpp,
  weight_params = list(a = 1.5, b = 0.5, prune = TRUE))

# Between-Trial Non-Monotonicity - Example 1
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(0, 1, 5, 6), lambda = 0.97, weight_fun = weights_mml)
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(0, 2, 5, 6), lambda = 0.97, weight_fun = weights_mml)
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(1, 3, 5, 6), lambda = 0.97, weight_fun = weights_mml)
# With Pruning
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(1, 3, 5, 6), lambda = 0.97, weight_fun = weights_mml,
  weight_params = list(prune = TRUE))

# Between-Trial Non-Monotonicity - Example 2
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(1, 5, 5, 5), lambda = 0.97,
  weight_fun = weights_mml)
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(2, 5, 5, 5), lambda = 0.97,
  weight_fun = weights_mml)
# With Pruning
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.15), n = 20,
  r = c(2, 5, 5, 5), lambda = 0.97,
  weight_fun = weights_mml, weight_params = list(prune = TRUE))

# Between-Trial Non-Monotonicity - Example 3
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.3), n = 20,
  lambda = 0.99, weight_fun = weights_cpp,
  weight_params = list(a = 2.5, b = 3), r = c(0, 0, 10, 10))
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.3), n = 20,
  lambda = 0.99, weight_fun = weights_cpp,
  weight_params = list(a = 2.5, b = 3), r = c(5, 7, 10, 10))
# With Pruning
basket_test(design = setupOneStageBasket(k = 4, p0 = 0.3), n = 20,
  lambda = 0.99, weight_fun = weights_cpp,
  weight_params = list(a = 2.5, b = 3, prune = TRUE),
  r = c(5, 7, 10, 10))

## Investigation of Monotonicity Conditions - Table 10
# Fujikawa's Design - Without Pruning
```

```r
for (i in 4:6) {
  cat("K =", i, "Within-Trial")
  print(check_mon_within(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = seq(0.5, 3, by = 0.5),
      tau = seq(0, 0.5, by = 0.1)
    ),
    details = FALSE
  ))

  cat("K =", i, "Between-Trials")
  print(check_mon_between(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = seq(0.5, 3, by = 0.5),
      tau = seq(0, 0.5, by = 0.1)
    ),
    details = FALSE
  ))
}

# Power Prior Design with CPP weights - Without Pruning
for (i in 4:6) {
  cat("K =", i, "Within-Trial")
  print(check_mon_within(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(0.5, 3, by = 0.5),
      b = seq(0.5, 3, by = 0.5)
    ),
    details = FALSE
  ))

  cat("K =", i, "Between-Trials")
  print(check_mon_between(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(0.5, 3, by = 0.5),
      b = seq(0.5, 3, by = 0.5)
    ),
    details = FALSE
  ))
}

```

```
# Fujikawa's Design - With Pruning
for (i in 4:6) {
  cat("K =", i, "Within-Trial")
  print(check_mon_within(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = seq(0.5, 3, by = 0.5),
      tau = seq(0, 0.5, by = 0.1),
      prune = TRUE
    ),
    details = FALSE
  ))

  cat("K =", i, "Between-Trials")
  print(check_mon_between(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = seq(0.5, 3, by = 0.5),
      tau = seq(0, 0.5, by = 0.1),
      prune = TRUE
    ),
    details = FALSE
  ))
}

# Power Prior Design with CPP Weights - With Pruning
for (i in 4:6) {
  cat("K =", i, "Within-Trial")
  print(check_mon_within(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(0.5, 3, by = 0.5),
      b = seq(0.5, 3, by = 0.5),
      prune = TRUE
    ),
    details = FALSE,
  ))

  cat("K =", i, "Between-Trials")
  print(check_mon_between(
    design = setupOneStageBasket(k = i, p0 = 0.15),
    n = 20, lambda = 0.99,
    weight_fun = weights_cpp,
    weight_params = list(
      a = seq(0.5, 3, by = 0.5),
      b = seq(0.5, 3, by = 0.5),
      prune = TRUE
```

```
    ),
    details = FALSE
  ))
}

# Influence of Pruning - Table 11
res_prune_fujikawa <- baskexact::opt_design(
  design = design_exact,
  n = 20, alpha = 0.05,
  weight_fun = weights_fujikawa,
  weight_params = list(
    epsilon = seq(from = 0.5, to = 3, by = 0.5),
    tau = seq(from = 0, to = 0.5, by = 0.1),
    prune = TRUE
  ),
  scenarios = scenarios,
  prec_digits = 3
)

res_prune_cpp <- baskexact::opt_design(
  design = design_exact,
  n = 20, alpha = 0.05,
  weight_fun = weights_cpp,
  weight_params = list(
    a = seq(from = 0.5, to = 3, by = 0.5),
    b = seq(from = 0.5, to = 3, by = 0.5),
    prune = TRUE
  ),
  scenarios = scenarios,
  prec_digits = 3
)

# Rejection Rates with Pruning - Table 12
params_prune <- list(
  "fujikawa" = list(
    lambda = res_prune_fujikawa$Lambda[1],
    weight_fun = weights_fujikawa,
    weight_params = list(
      epsilon = res_prune_fujikawa$epsilon[1],
      tau = res_prune_fujikawa$tau[1],
      prune = TRUE
    )
  ),
  "cpp" = list(
    lambda = res_prune_cpp$Lambda[1],
    weight_fun = weights_cpp,
    weight_params = list(
      a = res_prune_cpp$a[1],
      b = res_prune_cpp$b[1],
      prune = TRUE
    )
  )
)
```

```
924
925 methods_prune <- c("fujikawa", "cpp")
926 toer_prune <- list()
927 for (i in methods_prune) {
928   toer_prune[[i]] <- t(sapply(scenarios_toer, function(x)
929     unlist(unname(baskexact::toer(
930       design = design_exact,
931       p1 = x, n = 20,
932       lambda = params_prune[[i]]$lambda,
933       weight_fun = params_prune[[i]]$weight_fun,
934       weight_params = params_prune[[i]]$weight_params,
935       results = "group"
936     )))))
937 }
938
939 pow_prune <- list()
940 for (i in methods_prune) {
941   pow_prune[[i]] <- t(sapply(scenarios_pow, function(x)
942     unlist(unname(baskexact::pow(
943       design = design_exact,
944       p1 = x, n = 20,
945       lambda = params_prune[[i]]$lambda,
946       weight_fun = params_prune[[i]]$weight_fun,
947       weight_params = params_prune[[i]]$weight_params,
948       results = "group"
949     )))))
950 }
```

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr. Meinhard Kieser for proposing the topic of my thesis and for his constant support. I am very grateful for his guidance but also for the freedom that I had while working on my thesis.

I would also like to thank Dr. Marietta Kirchner for her helpful comments on the first draft of this thesis.

Furthermore, I would like to thank Alexander Ritz, Lukas Sauer and Paul Thalmann who were always open to discuss my newest ideas.

In addition, I thank all my colleagues at the Institute of Medical Biometry for the best working environment that one could ask for.

Finally, I would like to thank my wife Natalie for always supporting me, even when times were hard. I could not have done it without her.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema *A Computationally Efficient Basket Trial Design Based on Power Priors* handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift