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

Systematic approach and investigation of the statistical tools in basket trial designs

Inauguraldissertation
zur Erlangung des
Doctor scientiarum humanarum
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Moritz Pohl
aus
Esslingen am Neckar

2021

Dekan:	Prof. Dr. med. Hans-Georg Kräusslich
Doktorvater:	Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations and Symbols

$_{ij}$	indices for further specifications (e.g. stage, basket, path, ...)
$\mathbb{1}_{\{.\}}$	indicator function
$\mathcal{A}$	set of values with specified characteristics
α	significance level (type 1 error rate)
A, B	certain unspecified events
(a, b), (α, β)	shape parameters of beta distribution
$\mathcal{B}$	posterior probability with uniform prior
$B(.,.)$	beta function with respective input tupel
BB	(independent) beta-binomial
BHM	Bayesian hierarchical model based on Berry et al. (2013)
c	place holder for constant number (also with indices)
CPP	commensurate predictive prior
$\mathcal{D}$	observed data
d_{ij}	distance or divergence measure between basket i and j
δ	absolute difference
ESS	(prior) effective sample size
$\mathcal{E}$	non-specified arbitrary distribution
$\mathcal{F}$	non-specified arbitrary distribution, also with indices
FWER	family-wise error rate
$\mathcal{G}$	distribution or a class of distributions
γ	parameter in hierarchical model
$\Gamma(.)$	gamma function

H_0, H_1	null and alternative hypothesis
HBB	hierarchical beta-binomial model
i	basket number
I	total number of baskets
i^*	modified number of baskets
κ	multiplicative factor or hyperparameter
λ	probability for the choice of a distribution
MCMC	Markov Chain Monte Carlo
MLE	maximum likelihood estimator/estimation
MPP	marginal predictive prior
μ	parameter for expected value
n	number of observations/patients
ν	expected value in normal distribution or hyperparameter
$\mathbf{\Omega}$	matrix
ω	weight parameter
$\mathcal{P}$	p-value of one-sided Binomial test written as incomplete beta function
P[.]	probability function
p	response/event rate
p_0	reference response rate
ϕ	general tuning paramater or hyperparameter
PPP	posterior predicitive probabilty
r	number of responses, also with index for specific basket
r_{min}	minimum number of responses
R	random variable of number of responses
ϱ	hyperparameter
σ^2	variance of a normal distribution
$T1E$	type 1 error
$\mathcal{T}$	test statistic
τ, τ^*	threshold value and modified threshold value
τ^2	variance of a normal distribution
θ	logit-transformed response rate

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Advances in biological and medical research in the last decades have led to a shift in the treatment philosophy in the field of oncology. While the approach "one treatment for one disease" was common in the past, the focus has changed to more personalized and patient-centered treatment approaches. This new perspective has been enabled by two key advances, the first being the scientific, technical, and economical improvements in gene sequencing and the corresponding knowledge about the influence of specific genes on human health. The second advance inducing this new perspective is the development of treatments which directly take advantage of a known patient characteristic, e.g. a genetic aberration which is used for the treatment pathway (Kalia, 2013). Hence, the development of new treatments does not solely focus on one disease anymore, but aims to treat many diseases within a small subset of patients, who all express the same targeted characteristic. A direct consequence of these advances is the need for new trial concepts to investigate the efficacy of these targeted treatments (Berry, 2015). These trial concepts are gathered under the label 'master protocols' and contain trial designs which are called basket trial, umbrella trial, and platform trial. The literature has been diffuse on the definition of each of these trial designs such that no common nomenclature has been available. Woodcock and LaVange (2017) proposed to define a basket trial as a trial in which one treatment is investigated in multiple diseases or subtypes. They

defined umbrella trials as studies with multiple treatments for one single disease and platform trials as dynamic extensions of basket or umbrella trials. In a platform trial, baskets can be added or withdrawn during an ongoing protocol and, similarly, treatments can be added or withdrawn during the trial.

Practical trial examples for each of these three designs are available in the literature. The BATTLE-1 trial, presented by Kim et al. (2011), is an umbrella trial and investigated four different treatments in patients with chemotherapy-refractory non-small cell lung cancer. The I-SPY 2 trial (Barker et al., 2009) is an example for a platform trial. It investigates multiple treatments for the therapy of locally advanced breast cancer among ten different groups which were defined by three biomarkers. The trial is still ongoing and keeps on investigating potential treatments. A practical example for a basket trial is the BRAF V600 trial by Hyman et al. (2015). The authors investigated the treatment with vemurafenib among cancer patients who express a BRAF V600 positive-mutation. The patients were recruited to one of six prespecified cancer groups (e.g. colorectal cancer, cholangiocarcinoma), while a seventh group contained all other cancer types. The prespecified cancer groups contained 7 to 27 patients at the final analysis. The primary endpoint was overall response after 8 weeks. The overall response rate was analysed independently in each cancer group using the adaptive two-stage design by Simon (1989). The general conclusion of the trial was that the conduct of a basket trial is feasible. The medical results showed promising response rates in some baskets (e.g. non–small-cell lung cancer) but also non-promising results in colorectal cancer. So, even though the same genetic aberration was prevalent in all patients, the treatment showed different responses among the groups.

The trial design in Hyman et al. (2015) was a parallel arrangement of independent Simon two-stage designs, which could be considered as many individual trials under the shield of one common protocol. Since then, statistical researchers have proposed many new trial designs and statistical tools for the conduct and the analysis of basket trials. These designs also account for the idea to share information among the groups. The sharing is based on the assumption that the groups respond similarly due to the common genetic predisposition and the targeted treatment for this common characteristic. The statistical research has led to a variety of theoretical trial designs and tools for basket trials. The strong interest in the field

of master protocols has been reported by Park et al. (2019) and Meyer et al. (2020). Both of them performed systematic literature reviews and proved a high activity in master protocol research throughout the last decade. The number of publications shows an exponential growth where the peak has presumably not been reached yet. Also both reviews show that basket trials are the dominant trial design among the master protocols with respect to practical application and to methodological research. Consequently, when the work on this thesis started, the status quo of (statistical) basket trial designs was very dynamic, with a constant supply of new trial designs and statistical tools with increasing complexity.

1.2 Aim and structure of this thesis

The aim of this thesis is to investigate the basket trial design in its general structure as well as its specific statistical tools and techniques. The investigation of the general structure aims to define a systematic approach to the construction of basket trials. This includes a modular construction kit, together with an overview of available statistical methods and tools which are currently proposed for the use in basket trials, and also the elaboration of connections among the tools. In addition to the general investigation of basket trial designs, the individual statistical tools are investigated. The use of frequentist and Bayesian tools for decision making in basket trials, e.g. at an interim assessment, are analysed with the aim to evaluate differences or even equalities between the two statistical techniques. The sharing tool will be investigated regarding the feasibility of a non-transformed hierarchical model and it will be compared via simulations to the currently used basic hierarchical model with respect to their sharing property in the setting of basket trials. The overall motivation in all three aspects is to facilitate the accessibility to basket trials and to consolidate the statistical tools in order to increase the technical understanding of basket trials and to ultimately empower the practical application of basket trials in medical research.

This thesis is structured as follows. In Chapter 2, the methods are provided, introducing the methodological tools and the required knowledge needed for the elaboration of the results. The results are presented in Chapter 3 which consists of three sections. Section 3.1 presents the results for a systematic approach to construct a basket trial in a modular fashion, including an ordered presentation of available statistical tools in basket trials. The investigation

of the analytical connections between Bayesian and frequentist decision tools are presented in Section 3.2. Section 3.3 provides the results of the feasibility investigation for a non-transformed hierarchical model including its sharing properties in comparison to the current basic hierarchical model. In Chapter 4, the results are discussed together with their contribution to research as well as limitations and directions for further research. In Chapter 5, the thesis is summarised, once in English and once in a literal translation to German. Additional results are presented in Appendix A, while Appendix B contains relevant R program code for this thesis.

Chapter 2

Methods

2.1 Basket trials

2.1.1 Concept of basket trials

Parts of this Subsection 2.1.1 are already published in the article *Categories, components, and techniques in a modular construction of basket trials for application and further research* by Pohl et al. (2021). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

Basket trials are clinical trial designs which have emerged in the last decade with the rise of personalized medicine. The knowledge about genome sequencing has led to a different view on the categorisation of cancer types, instead of the localization from where the cancer origins, the focus shifted to the genetic predisposition of the cancer. Clinical research therefore started to focus on treatments which take advantage of characteristics that are associated with the genetic predisposition of the cancer. The guiding assumption is that the treatment has similar response among cancer types with a common genetic predisposition irrespective of the localization (cf. Redig and Janne, 2015). As for all treatments, these targeted therapies also have to prove their treatment effect in clinical trials, and basket trial designs were developed for this purpose. An official definition of basket trials does not exist. However, Woodcock and LaVange (2017) suggested that basket trials are studies with one treatment for multiple

diseases or subtypes. The wording in literature for the diseases or subtypes is not consistent and various names are used. These include subpopulations, indications or strata, but they are also called baskets (e.g. in Cunanan et al., 2017b; Chu and Yuan, 2018a; Psioda et al., 2019). The latter wording is used in this work, meaning that a basket trial consists of several baskets, with each basket representing a disease or subtype. The separation of the subtypes into different baskets covers, e.g. the different localizations in the light of potentially different treatment effects even though they all share the targeted genetic predisposition. Initially, basket trials were developed for an oncological setting, but they are also of interest in other medical fields, like for example in psychiatry (cf. Joshi and Light, 2018).

2.1.2 Basket trial design of Cunanan et al. (2017b)

The goal of the basket trial design proposed by Cunanan et al. (2017b) is to investigate whether the treatment works in general, but also in which particular baskets the binary response rate is promising. The starting point are independent Simon two-stage designs (see Simon, 1989), where each disease is investigated in a separate clinical trial. The design of Cunanan et al. (2017b) aims to improve the efficiency of this approach when several indications are investigated in parallel using one common treatment. The core innovation of the proposed design is to evaluate the basket-wise response at the interim assessment and to decide subsequently whether the baskets continue independently in a heterogeneous path or if all baskets are combined into one group in the homogeneous path. The premise of Cunanan et al. (2017b) is that the option to combine baskets in the second stage leads to a higher power (or the same power with less patients) to declare the treatment efficacious compared to the individual Simon two-stage trials. The benefits of the combination must be considered together with power losses in the evaluation of basket-individual effects. The design is used in Section 3.2 and, therefore, here introduced in detail.

The design of Cunanan et al. (2017b) is a two-stage basket trial and its schematic composition of decision nodes is depicted in Figure 2.1. In stage 1, it is equivalent to the first stage of an adaptive Simon two-stage design. The first modification is the heterogeneity assessment among the baskets when the planned patient accrual of n_1 patients in stage 1 is completed. The heterogeneity is evaluated with an $i \times 2$ (Fisher) exact test with respect to H_0 : *homogenous response among all baskets*. The threshold for the test is a tuning pa-

rameter of the design and the authors explicitly state that its purpose is to guarantee the desired design characteristics of the complete trial design and not to interpret the test as an autonomous decision of heterogeneity. The exact test determines the path for the succeeding steps in the trial. In the heterogeneous path, each basket is further evaluated individually. This is similar to the approach in the Simon two-stage design. When pursuing the homogeneous path, all baskets are pooled together into one group. After the heterogeneity assessment, the directly following trial node is the futility assessment. The futility assessment requires a minimum number of responses to continue to stage 2. In the heterogeneous path, the minimum number of responses r_s refers to each basket individually and a basket continues if $r_i \geq r_s$, otherwise basket i is stopped due to futility. The decisions in the homogeneous path are made in the same way, however, the minimum number of responses r_c is compared to the total number of responses accross all baskets. In stage 2, additional patients are recruited, either n_{2i} patients to each of the $i^* \leq i$ baskets that passed the futility bar in the heterogeneous path, or n_2 patients to the combined group in the homogeneous path. The parameters n_1, r_s, r_c, n_{2i}, and n_2 are tuning parameters. The final analysis is conducted using the one-sided exact binomial test. The corresponding null hypothesis is that the response rate is below or equal to the null value p_0, and the alternative hypothesis is that the response rate exceeds p_0. In the heterogeneous path each basket is evaluated individually. The significance level α_s is a tuning parameter and is adjusted with a Bonferroni correction $\frac{\alpha_s}{i^*}$ by the number of baskets in stage 2. In the homogeneous path, all baskets are evaluated together in the combined set and one binomial test is conducted to evaluate if the treatment works in all baskets or not. The respective significance level α_c is a tuning parameter.

The choice of the tuning parameters aims to control the overall false positive rate at a prespecified level. Three metrics are used by Cunanan et al. (2017b) to evaluate this. First, the family-wise error rate (FWER) which is defined as the probability to declare at least one basket promising when in truth all baskets have a response rate equal to the null value p_0. The second metric is the marginal power which describes the probability to declare efficacy for a truly efficacious basket. This also allows to evaluate the basket-wise type 1 error (T1E), which is the probability to reject the null hypothesis when in truth the treatment does not work in this basket. The third metric is the expected sample size for the trial. The optimized design controls the family-wise error and has an optimal trade-off between

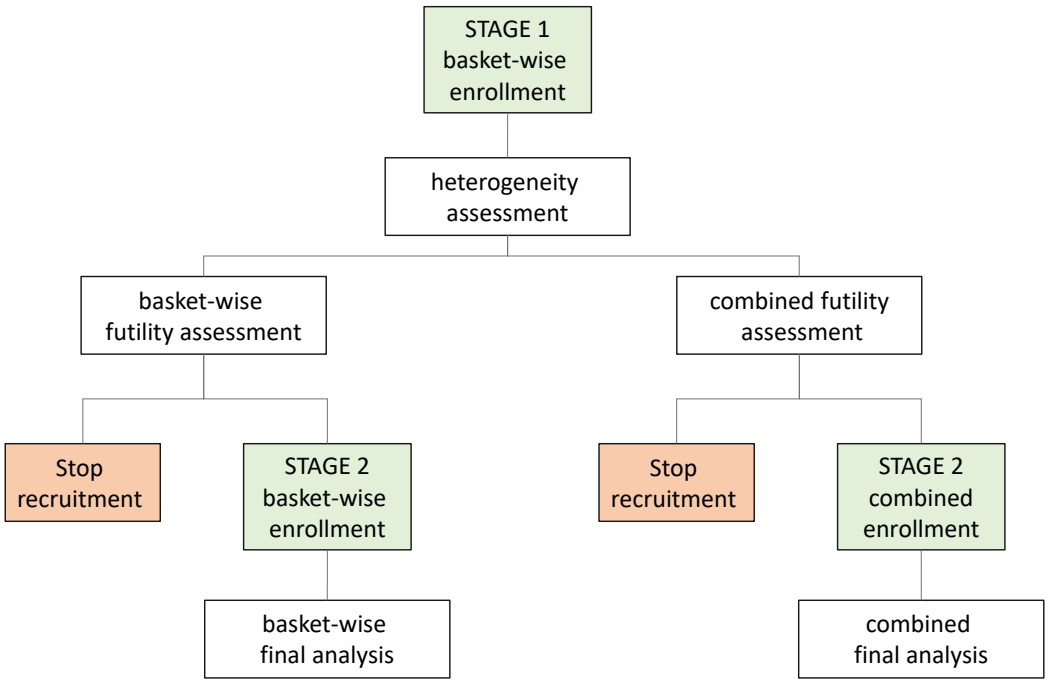


Figure 2.1: *Basket trial design of Cunanan et al. (2017b) with decision nodes of the design according to the flow chart in the original publication.*

power and expected sample size. The tuning parameters that fulfill these requirements are determined via simulations. The authors assign specific values to n_1, n_2, r_s, and r_c. They do so because of the high dimension of the tuning parameters and the resulting extensive computational burden when the optimization accounts for all parameters. The assigned values reflect logical arguments and previous assumptions about the trial purpose. For example, Cunanan et al. (2017b) propose to use $r_s := 1$, because then, baskets without any observed response are stopped. Similarly, for r_c at least I responses are demanded, where I is the number of baskets in the trial.

Moreover, the authors propose minimum and maximum restrictions for the number of patients per basket. These rules are trial specific and can be individually chosen for every decision node. The restrictions shall prevent a too high influence of individual baskets, especially in the case of unequal accrual.

2.1.3 Binomial test in general and for basket trials

The binomial test is used to evaluate the probability of a certain event in comparison to an assumed probability value (Agresti, 2007). In clinical trials, the probability can represent a response rate which describes how likely a patient is to respond to a certain treatment. The probability is denoted by p. The two-sided hypothesis is

$$H_0 : p = p_0 \qquad H_1 : p \neq p_0$$

with a null or reference value p_0 with which the observed outcomes are compared. In the one-sided test setting the hypothesis is given by

$$H_0 : p \leq p_0 \qquad H_1 : p > p_0$$

and the one-sided direction can also be the other way around.

The underlying observed data to evaluate the hypothesis consists of the number of observations n and the number of events (e.g. responses, successes). For each individual observation, the outcome is binary. The total number of events is denoted by r and is a realisation of a random variable R following a binomial distribution $Bin(n, p)$. The test statistic contains the observed data and is used to evaluate the hypothesis. For the binomial test, the test statistic is r given the number of observations n. This test statistic has a binomial distribution under the assumption that the null hypothesis H_0 is true. The decision with respect to the hypothesis can be made based on the comparison of the p-value to the significance level α, or based on the comparison of the test statistic with its critical values. The critical values are those realisations of r which change the decision from staying with the null hypothesis to rejecting it. In the two-sided setting, the critical values are a tuple of a lower boundary and an upper boundary that represent number of events. In the one-sided setting the critical value is either the minimum or the maximum number of events, this depends on the direction of the alternative hypothesis. For the notation above, the critical value is the minimum number of events.

The binomial distribution of the test statistic results in a p-value that consists of the sum over all probabilities of the test statistic outcomes that are equal or less likely than the observed

r events under the assumption that H_0 is true. This set of outcomes represents the scenarios that are at least equally extreme as the observation and is defined as $\mathcal{A} := \{k : \mathrm{P}[R = k|n, p_0] \leq \mathrm{P}[R = r|n, p_0]\}$. The respective p-value for the two-sided test is then given by

$$\mathrm{P}[R \in \mathcal{A}|n, p_0] = \sum_{x \in \mathcal{A}} \mathrm{P}[R = x|n, p_0] = \sum_{x \in \mathcal{A}} \binom{n}{x} p_0^x (1 - p_0)^{n-x}.$$

In the one-sided setting, where the alternative hypothesis is as above, the p-value is given by

$$\mathrm{P}[R \geq r|n, p_0] = \sum_{x=r}^{n} \mathrm{P}[R = x|n, p_0] = \sum_{x=r}^{n} \binom{n}{x} p_0^x (1 - p_0)^{n-x}$$

where the sum goes over the observed and more extreme outcomes of the test statistic. The extremeness refers to the direction of the alternative hypothesis under the assumption that H_0 is true.

The Central Limit Theorem allows to approximate the distribution of the test statistic to a standard normal distribution and consequently statistical tests based on a normal distribution are possible. The approximate tests require a larger number of observations n (cf. Fahrmeir et al. (2007)). In clinical trials of early phases, the number of observations are often modest, then the exact test based on the binomial distribution should be applied (cf. Sebastiao and St. Peter (2018), Agresti (2007)). Especially in basket trials which investigate a rare disease, the number of observations does not justify an approximate test. The exact binomial test is a frequentist tool to evaluate an observed binary treatment response against a prespecified control or reference value p_0. The test can therefore be applied in basket trials without control group to assess whether the treatment is promising at interim stages and for the final analysis to make informed decisions (see e.g. Cunanan et al., 2017b).

2.2 Bayesian statistics in basket trials

This section covers the Bayesian methodologies that are used in the context of basket trials relevant for this thesis. This requires a general approach towards Bayesian statistics including conjugate models as well as hierarchical Bayesian models. Moreover, known and for this thesis relevant connections between Bayesian and frequentist statistics are introduced.

2.2.1 General methodologies of Bayesian statistics

The source of Bayesian statistics is Bayes' theorem. The theorem was defined by Thomas Bayes and was posthumously published as *An essay towards solving a problem in the doctrine of chance* by Bayes and Price (1763).

Let A and B be two different sets of events that take place with non-zero probability. Then Bayes' theorem is defined as

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A] \cdot \mathrm{P}[A]}{\mathrm{P}[B]}. \tag{2.1}$$

This idea can be transferred to density functions. The probabilities of certain (conditional) events are then replaced by densities of parameters. In this section, θ represents an unknown and continuous random variable with a distribution on the parameter space Θ. The observation x is a realisation of the random variable X and Bayes' theorem results in

$$f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{f(x)}. \tag{2.2}$$

The density of x under the assumption that θ is a distribution parameter of X is given by $f(x|\theta)$ and is called the *likelihood* of the observation. The wording *data* and notation $\mathcal{D}$ are interchangeable for the observation x, since the data set $\mathcal{D}$ consists of the observations. The density function $f(\theta)$ is called the *prior* density. The prior represents the initial knowledge or assumptions about the distribution of parameter θ. The resulting conditional distribution of $\theta|x$ with corresponding density function $f(\theta|x)$ is called the *posterior distribution*. The density in the denominator $f(x)$ is called the *marginal likelihood* and is the unconditional density of the observation x. Due to the law of total probability the marginal likelihood can be determined by the likelihood and the prior

$$f(x) = \int_{\Theta} f(x|\theta) f(\theta) d\theta.$$

In case of a discrete parameter space, the integral is replaced by a sum. The unconditional density $f(x)$ in the denominator is a constant value. Hence, the posterior is proportional to the likelihood multiplied with the prior. This reduces the formula for the density of the posterior distribution to

$$f(\theta|x) \propto f(x|\theta) \cdot f(\theta). \tag{2.3}$$

The property of the marginal likelihood is to normalize the posterior distribution such that it fulfills the requirements for a valid density function ($\int_\Theta f(\theta|x)d\theta = 1$). In general, the dimension of x and θ is arbitrary and can be multidimensional. With higher dimensions, the complexity to calculate the posterior distribution increases.

The core element of Bayesian statistics is the posterior distribution. It represents the updated distribution of the parameter θ after observing the data, which means the posterior distribution is an updated prior distribution. Hence, in the Bayesian approach the parameter θ is considered random, whereas in frequentist statistics the observation is considered random with a fixed value for the parameter. The choice of the prior distribution is crucial because together with the likelihood distribution it determines the form of the posterior distribution. Relevant aspects in the prior choice are the amount of information that should be contained in the prior and the form of the distribution. The latter means that the prior distribution can come from known distribution families (e.g. normal, beta), can be a mixture of distributions, or can also take non-parametric forms. Wise choices of prior distributions in combination with the known distribution of the data can have analytical advantages and these are introduced in the following subsection.

2.2.2 Conjugate models

A conjugate Bayesian model consists of a prior, likelihood combination that results in a posterior distribution from the same class as the prior. In formal notation a class of distributions $\mathcal{G}$ is called conjugate with respect to the likelihood of the observed data $f(x|\theta)$, if the posterior distribution $f(\theta|x)$ is in $\mathcal{G}$ for every observation x and any prior distribution $f(\theta) \in \mathcal{G}$. The most trivial choice for $\mathcal{G} := \{\text{all distributions}\}$ is of limited practical benefit (Held and Bové, 2020). A more restrictive definition of class $\mathcal{G}$ can facilitate the analytical calculation of the posterior probability, as for example when $\mathcal{G}$ only consists of one family of distributions, because then the posterior distribution with respect to a certain likelihood only differs from the prior distribution in the parameters. Gelman et al. (2004) call that case a natural conjugate combination. Throughout this work conjugate prior, likelihood combinations are considered as natural ones. A classic example for a conjugate model is the normal-normal scenario where the likelihood of the data is given by $X|\mu \sim N(\mu, \sigma^2)$ with σ^2 known and a normal prior distribution for parameter $\mu \sim N(\mu_\mu, \sigma_\mu^2)$. The resulting posterior distribution

is again a normal distribution with a weighted mean of the prior and the data, and a variance that is smaller than σ^2 and σ_μ^2.

The beta distribution is a conjugate prior distribution for data that has a binomial distribution $Bin(n, p)$. For the beta-binomial model, the observed data is denoted as r, a realisation of the random variable R. Let the prior for the response rate p be given by $Beta(a, b)$. The posterior distribution is derived according to Bayes' theorem in Equation 2.3

$$\begin{aligned} f(p|r) &\propto f(r|p) \cdot f(p) \\ &= \binom{n}{r} p^r (1-p)^{n-r} \cdot \frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1} \\ &\propto p^{r+a-1} (1-p)^{n-r+b-1} \end{aligned}$$

and hence for the posterior distribution it follows

$$p|r \sim Beta(a + r, b + n - r).$$

Both examples demonstrate that the observed data is additional information that updates the prior distribution by modifying the distribution parameters. In application, this characteristic is a practical advantage because the conjugate models facilitate the understanding and the calculations for the posterior (Gelman et al., 2004). In general, for any likelihood with a distribution that fulfills the requirements of an *exponential family*, there exists a natural conjugate prior distribution. For further specification see page 41-42 in Gelman et al. (2004).

Non-conjugate combinations are more dominant in practical applications, because there is not always a prior-likelihood combination that reflects the underlying problem. Additionally, conjugate models become more challenging or are mostly impossible in complex structures, like for example in hierarchical models (cf. Subsection 2.2.3). The posterior distribution in non-conjugate models is derived in numerical Markov Chain Monte Carlo (MCMC) sampling using different numerical algorithms.

2.2.3 Hierarchical models

A hierarchical model adds another level to the combination of likelihood and prior. Instead of fixed parameters, the prior distribution is defined by parameters which are random by themselves. The distribution of these parameters, here denoted as γ, is called hyperprior distribution and the respective distribution parameters are called hyperparameters (Christensen et al., 2011).

The additional level allows more variability among the investigated units. In a basket trial each basket represents one unit. If the prior distribution is fixed, then all basket-wise posterior distributions are based on this prior distribution alone and each basket is evaluated independent of the others. However, in a hierarchical model the additional level allows many different prior distributions via the hyperprior. When applying a hierarchical model in a basket trial, all basket-individual parameters are modeled from one common distribution which has its origin in the hyperprior distribution (Gelman et al., 2004). The latter reflects the underlying *exchangeability* assumption in hierarchical models, which formally means that the joint distribution of all basket-individual parameters $f(\theta_1, ..., \theta_I)$ is invariant to any permutation in the basket index $i \in \{1, ..., I\}$.

Since the basket individual parameters are assumed to come from the same common distribution, the variance of the prior distribution determines how similar the baskets are. A prior distribution with low variance means that all baskets are similar whereas a high variance allows rather different values and, therefore, dissimilarities between the baskets. Because the parameters of the prior are random, they are adapted in the sense of Bayes' theorem. The additional level ensures that each basket distribution is guided by the same distribution, allows more variability between baskets, and enables to share information between baskets. In formal notation the hierarchical model can be given as

Likelihood	$x\|\theta, \gamma \sim f(x\|\theta, \gamma)$
Prior	$\theta\|\gamma \sim f(\theta\|\gamma)$
Hyperprior	$\gamma \sim f(\gamma).$

The parameters θ and γ can also be vectors in the case of multidimensional distributions and the likelihood describes unit-wise observations. The joint posterior distribution of the parameters is then given by

$$\begin{aligned} f(\theta, \gamma|x) &= f(\theta|\gamma, x) \cdot f(\gamma|x) \\ &\propto f(x|\theta, \gamma) \cdot f(\theta|\gamma) \cdot f(\gamma) \end{aligned}$$

as a multiplication of the conditional distribution $f(\theta|\gamma, x)$ for the parameter θ based on the hyperparameter and the observed data, with the marginal distribution $f(\gamma|x)$ for the hyperparameter γ based on the observed data. The likelihood depends on both the prior and the hyperprior, however, the dependence on the hyperprior is via the dependence of the prior on the hyperprior and consequently one could reduce $f(x|\theta, \gamma)$ to $f(x|\theta)$. In a hierarchical Bayesian model, the posterior distribution depends on the distribution of the likelihood, the prior, and the hyperprior.

In the previous subsection the conjugate beta-binomial model was introduced and this model is now extended to a hierarchical beta-binomial model. The hierarchical binomial model can be depicted as

$$\begin{aligned} r|p &\sim Bin(n, p) \\ p|(a, b) &\sim Beta(a, b) \\ (a, b) &\sim \mathcal{F} \end{aligned}$$

with a hyperdistribution $\mathcal{F}$. Apart from a multidimensional distribution for the hyperprior tupel (a, b) one can also define individual and independent hyperdistributions for each element $a \sim \mathcal{F}_1$ and $b \sim \mathcal{F}_2$. The random values of a and b then determine the prior distribution of the response parameter p which is needed to describe the distribution of the observed data. Instead of distributions for a and b, one can use the characteristics of the beta distribution to improve the interpretation of the hyperprior level by setting distributions on the expected value $\mathbb{E}[X] = \frac{a}{a+b}$ and on a measure that represents the variance of the distribution. The variance of a beta distribution is given by $Var(X) = \frac{ab}{(a+b+1)(a+b)^2}$. The sum $a + b$ is part of the variance and moreover is the effective sample size (ESS) of a beta distribution

(Morita et al., 2008). The ESS represents the number of observations that are included in a prior. Hence, the ESS, given by $a+b$, can be used as an easily interpretable value to describe the width of the prior beta distribution (Christensen et al., 2011).

The hyperparameter constitutes the highest level and creates a common link between all elements on the lower levels. Information among elements on the lower levels can be shared through this link. This property can be used in basket trials, and a basket trial design with a Bayesian hierarchical model is introduced in the next subsection.

2.2.4 Bayesian hierarchical basket trial design of Berry et al. (2013)

Berry et al. (2013) proposed to use a Bayesian hierarchical model in a clinical trial where several groups of patients are all treated in the same way. Their intention is to borrow information among the baskets regarding the primary outcome, the binary tumor response. Hence, the treatment effect is displayed as basket-individual response rates, denoted as p_i for basket i. The authors define a null response rate p_0 and define a target response rate p_1 as promising. The authors transform the basket-individual response rate p_i to a continuous scale

$$\theta_i := log\left(\frac{p_i}{1-p_i}\right) - log\left(\frac{p_1}{1-p_1}\right)$$

where the last element is a constant value. The sharing of information between the baskets is conducted with the following Bayesian hierarchical model

$$\begin{aligned} \theta_i &\sim N(\mu, \sigma^2) \\ \mu &\sim N(-1.34, 10^2) \qquad \forall i = 1, ..., I. \\ \sigma^2 &\sim IG(0.0005, 0.000005) \end{aligned} \tag{2.4}$$

The expected value μ determines the center around which the θ_i's and, respectively, the p_i's are located. The variance σ^2 determines how large the range of values for the θ_i's is. Large values for σ^2 mean that rather different values are likely, whereas small σ^2 values restrict the value range. Consequently, the variance determines the amount of sharing between the baskets. The variables μ and σ^2 are both random variables, therefore, according to the Bayesian idea, their distributions are adapted based on the observed data in all baskets.

The distributions and their hyperparameters for μ and σ^2 in Equation 2.4 were specified by the authors. The expected value for the distribution of μ reflects the null scenario with $p_0 = 0.10$ and an assumed target response of $p_1 = 0.30$. The authors explicitly also allow other parametric distributions, e.g. uniform or half-Cauchy, as hyperprior distributions for σ^2.

The authors allow adaptive decisions during the trial which includes stopping of baskets due to futility and stopping for efficacy. These interim assessments are conducted after in total 10 patients were observed, and then after every 5 additional patients. The decisions are taken on the posterior distributions of the basket-individual response rates that were calculated with the hierarchical model, meaning that information between the baskets is shared before interim decisions are made. The final analysis again uses the posterior distributions of the basket-individual response rates from the hierarchical model to declare baskets promising or not.

2.2.5 Connections of frequentist and Bayesian methodologies

The comparison and connection of frequentist and Bayesian methodologies with respect to a one-sided hypothesis has been investigated by Zaslavsky (2010). The author showed that there are prior choices for the beta-binomial model such that the posterior probability to not exceed the null response rate ($\mathrm{P}[p \leq p_0|\mathcal{D}]$) is either larger, equal or smaller than the respective p-value of a one-sided binomial test (order of hypothesis as in Section 2.1.3). For a uniform prior ($Beta(1,1)$), it was shown that the mentioned posterior probability is smaller than the respective p-value.

In frequentist statistics, the focus strongly lies on the type 1 error (T1E), whereas in Bayesian statistics it is sometimes neglected. Still, whenever a decision is made, there is the possibility for false decisions. The T1E is an important measure to quantify this, irrespective of whether a frequentist or a Bayesian tool is applied. The wording for type 1 error in this thesis is either T1E or false-positive rate.

2.3 Further relevant methods

2.3.1 Properties of the beta and the gamma function

The properties and connections among the beta and the gamma function are provided in many basic statistics books, e.g. in Chapter 39 of Arens et al. (2009).

The density of a beta distribution, $Beta(a, b)$, uses a beta function $B(a, b)$ as a normalizing constant. The beta function is defined as

$$B(a,b) := \int_0^1 t^{a-1}(t-1)^{b-1}dt$$

and can be rewritten as a quotient of gamma functions $\dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

The gamma function $\Gamma(a)$ is defined as $\int_0^\infty t^{a-1}e^{-t}dt$ and important characteristics of the gamma function are

$$\Gamma(1) = 1, \qquad \Gamma(a+1) = a \cdot \Gamma(a), \qquad \Gamma(n+1) = n!, \text{ for } n \in \mathbb{N}.$$

2.3.2 Logit-normal distribution

The characteristics of the logit-normal distribution are based on Frederic and Lad (2008) and on the original work on (logit) transformations of variables with normal distribution of Johnson (1949).

A random variable X has a logit-normal distribution if the logarithmic transformation of the odds of X is normally distributed meaning that from $X \sim logitN(\mu, \sigma^2)$ follows $logit(X) := log\left(\dfrac{X}{1-X}\right) \sim N(\mu, \sigma^2)$. The density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x(1-x)} \cdot exp\left(-\frac{1}{2}\left(\frac{logit(x)-\mu}{\sigma}\right)^2\right)$$

The expected value for X and its variance cannot be given in closed analytical form and numerical methods are required to calculate them. The median of X is given by $logit^{-1}(\mu) := expit(\mu) := \dfrac{exp(\mu)}{(1+exp(\mu))}$. The logit-normal distribution is not symmetric, except for $\mu = 0$.

In that case the median ($expit(0) = 0.5$) corresponds with the expected value and hence $\mathbb{E}[X] = 0.5$. The logit-normal distribution is defined on $(0, 1)$ and converges to 0 at the boundaries. For small values of the variance σ^2, the distribution is unimodal, for larger variances the distribution is bimodal with a U-shaped form. The derivative of the density function is 0 for all solutions of the equation $logit(x) = \sigma^2(2x - 1) + \mu$. The density has a bimodal form (two local maxima and one local minimum) if the solution consists of three values for x, and it is unimodal if only one solution exists.

2.3.3 Software

The calculations, including the simulations, and the graphs in this thesis were done with the statistical software R, version 3.5.1 and higher (R Core Team, 2021). The operating surface to write and run the program code was RStudio (RStudio Team, 2021). The basic R software was extended by several packages which were needed for this thesis. The most relevant one was the package rjags (Plummer, 2019) which was used to access the MCMC sampler JAGS (Plummer, 2003) which then calculated the posterior distributions and returned them to R. The package ggplot2 (Wickham, 2016) was used for the creation of the graphs. The R program code for the simulations is given in Appendix B.2.

Chapter 3

Results

This chapter presents the elaborated results of the investigations of basket trials for this thesis. This chapter is structured according to the aims of this thesis. The first section presents a systematic approach to construct basket trials in a modular fashion and an ordered presentation of available and potential tools that can be used in basket trials. The second section presents analytical connections between Bayesian and frequentist decision tools for interim and final assessments. The final section investigates the feasibility of a non-transformed hierarchical model to share information in basket trials including its sharing properties which are compared to the current basic hierarchical model in a simulation study.

3.1 Categorization and modular construction of basket trials

Parts of this Section 3.1 are already published in the article *Categories, components, and techniques in a modular construction of basket trials for application and further research* by Pohl et al. (2021). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

The highly dynamic research field of basket trials has led to an unordered accumulation of designs and statistical tools. Hence, an ordered approach to basket trials is elaborated with the intention to categorise basket trials, to introduce consistency to the presentation of the statistical methods and tools, and to design basket trials with a modular framework.

The categorization of existing basket trial designs is proposed to be performed on two metrics: firstly, the purpose of the trial and, secondly, the applied statistical techniques. Most basket trials aim to detect evidence for a possible treatment effect in an early stage of clinical development. Hence, they are mainly classified as phase II trials with the intention to generate data and evidence which can be used to plan a potential phase III trial. Jin et al. (2020a) proposed a design for early development which investigates whether the treatment works in at least one basket and explicitly named it a proof-of-concept (PoC) design. Other authors, like Berry et al. (2013), Neuenschwander et al. (2016), Simon et al. (2016), Chen and Lee (2019), Chu and Yuan (2018b), Chu and Yuan (2018a), Hobbs and Landin (2018), Psioda et al. (2019), Zheng and Wason (2020), Fujikawa et al. (2020), Zhou and Ji (2020), Chen and Lee (2020), Jin et al. (2020b), Lyu et al. (2020), and Asano and Hirakawa (2020) locate their designs to early clinical development in phase II aiming to either detect indications where the treatment works or at least identify promising efficacy results. The designs of Chen et al. (2016) and Li et al. (2017) move on to confirmatory phase III basket trials and intend to achieve approval for the treatment in multiple indications that are covered in one basket trial. Chen et al. (2016) pursue to control the type 1 error (T1E) and use frequentist statistics to achieve that. On the other hand, most of the proposed designs for early clinical phases apply Bayesian techniques to analyse and conduct the trial. However, among the early phase designs, there are also frequentist approaches (Cunanan et al., 2017b; Zhou et al., 2019), while Liu et al. (2017) combines both techniques. Consequently, the proposed categorization of basket trials uses the purpose (early phase exploratory, late phase confirmatory) of the trial and the applied statistical techniques (frequentist, Bayesian, combination of both). In Table 3.1, the so far available trial designs are categorised.

The evolution of basket trials has resulted in designs that reach from early phases up to confirmatory trials, containing a variety of statistical methods. The statistical methods all have in common, that they target the assumed connection between the administered treatment and the patients' characteristic in order to acquire as much information as possible from the observed data. The data are represented through the primary endpoint, for which the proposed basket trial designs address different scales. Still, the majority of the designs use a binary endpoint because this covers the frequently used treatment response as the endpoint in exploratory trials.

Table 3.1: *Proposed categorisation of available basket trial designs according to the two metrics: purpose (rows) and statistical technique (column).*

	Bayesian	**frequentist**	**both**
PoC	Jin et al. (2020a)		
phase II	Berry et al. (2013) Neuenschwander et al. (2016) Simon et al. (2016) Chen and Lee (2019) Chu and Yuan (2018b) Chu and Yuan (2018a) Hobbs and Landin (2018) Psioda et al. (2019) Zheng and Wason (2020) Fujikawa et al. (2020) Zhou and Ji (2020) Chen and Lee (2020) Jin et al. (2020b) Lyu et al. (2020) Asano and Hirakawa (2020)	Cunanan et al. (2017b) Zhou et al. (2019)	Liu et al. (2017)
phase III		Chen et al. (2016) Li et al. (2017)	

The proposed modular framework consists of four elementary design components of a basket trial. The first component is the sharing of information between baskets. It is the core element of a basket trial, is justified by the common patient characteristic, and therefore is an obligatory component in a basket trial design. The sharing of information intents to raise the power to detect individual baskets with clinically relevant treatment effects. The time points when sharing is applied can be prespecified to defined nodes of the trial, or can regularly take place before decisions regarding the treatment efficacy are made. The frequency of conducting sharing is a trial design characteristic, however, sharing must be applied at least once throughout a basket trial. The second component is the futility assessment. It prevents patients from treatment with futile treatments and additionally protects the sponsor's resources in form of budget and workforce. The interim futility assessments are optional, they can be conducted once or several times during a basket trial. The third component is the interim efficacy assessment. It terminates the recruitment of patients to baskets with convincing evidence in favour of the investigated treatment. In exploratory trials, an interim efficacy stop accelerates the subsequent initiation of a confirmatory trial, while the interim efficacy stop in a confirmatory basket trial might lead to an accelerated market approval. The

interim efficacy assessment is also optional, might be applied once or several times and can, but does not have to, be employed parallelly with the futility assessment. The final analysis of the baskets at the end of the trial is the fourth design component of a basket trial and is also obligatory. The final analysis investigates the baskets according to the objective of the trial. Each component in the modular approach for basket trials has its purpose. The sharing of information (first component) and the final analysis (fourth component) are the defining elements of a basket trial. The key element of a basket trial is the sharing of information because of its medical justification of similar treatment behaviour by the mutual predisposition among the baskets. The interim futility and efficacy assessments contribute to an efficient conduct of basket trials with respect to the resources (e.g. sample size), ethical aspects, and type 1 and 2 error rate control. The components per se are empty elements of the modular framework, and a basket trial becomes tangible when the components are orderly arranged in a workflow and filled with statistical tools such that the components fulfill their designated task.

The workflow is defined by the arrangement of the four components for which a multitude of combinations are possible. The only restrictions are to share information at least once and to have a final analysis. The majority of trial designs which were proposed in literature incorporate the sharing of information before the interim futility and efficacy assessments, and again before the final analysis. This reflects the intuitive strategy to use as much information as available for an informed decision. However, such an order of components is not a must, as for example in Liu et al. (2017), where the futility assessment is conducted before information is shared among the baskets. In general, the modular framework allows flexibility on how the components of a basket trial are arranged. The schematic workflow of a basket trial and the exemplary arrangement of the components is given in Figure 3.1.

The statistical tools for the components can be chosen according to the preferences of the research team and the purpose of the trial. The trial designs proposed in the literature apply different statistical tools and serve as a portfolio for the tools in basket trials that are designed with the modular framework. The modular basket trials can apply different tools from different designs. The flexibility of the tools allows basket trials which consist of

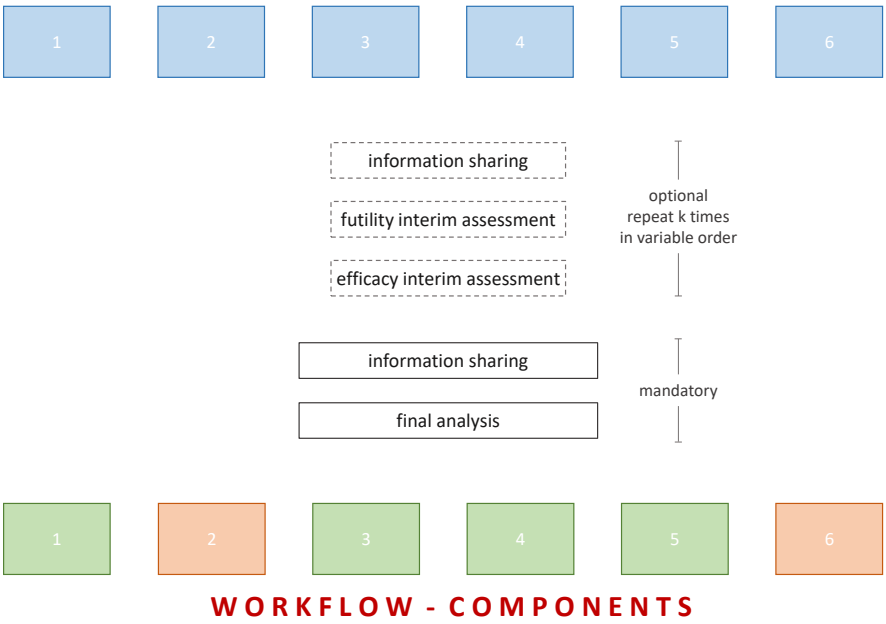


Figure 3.1: *Schematic display of a basket trial with $I = 6$ baskets (blue boxes), the arrangement of optional and mandatory components throughout the trial, and an exemplary presentation of promising (green) and non-promising (red) baskets after the final analysis. Adapted from Pohl et al. (2021).*

the sharing tool from one design, the interim (futility and/or efficacy) assessment tool from another, and the final analysis with a tool from a third design.

The modular framework for basket trials allows to design a new basket trial in a systematic way, it allows to arrange the four components in a workflow such that the trial fulfills practical requirements, and it allows to choose the statistical tools to analyse trial results adequately. The informed decision for the statistical tools is important in applied basket trials to correctly investigate the underlying medical research question. Consequently, the technical aspects of already published tools for each component are presented in the following subsections.

3.1.1 Consistent notation for basket trial tools

A systematic approach to the statistical tools for the components requires a structured and consistent notation throughout all tools. This notation is specified here before the statistical tools are investigated in detail.

An index i allocates a variable to basket i. Variables without an index represent the global parameter value that is valid for all baskets. Additional indices are explicitly defined in

advance if they are required. The variable n is the number of patients, r describes the number of responses for the binary outcome and p denotes the response rate. The so far observed data is represented by $\mathcal{D}$ and the explicit data from basket i is denoted by $\mathcal{D}_i := \{n_i, r_i\}$. Transformations of p are denoted as θ, and if not stated differently θ is defined as $\theta := \text{logit}(p) = \log(\frac{p}{1-p})$. In general, θ can also represent a continuous parameter with support on the real numbers. Null and alternative values of p and θ are indexed with 0 and 1, respectively. A pairwise distance measure between two baskets i and j is given by d_{ij}. Further details about the type of distance, or divergence measure are given in respective subsections. The variable λ denotes the probability of a distribution in a Bayesian mixture distribution. The variable ω describes a weight. The variable γ is a general parameter of a distribution. The variable τ describes a threshold value for decision making and the index F indicates interim futility assessment, index E stands for interim efficacy assessment and A indicates the final analysis at end of the trial. Additional indices in superscript distinguish parameters of the same type but with different values. The function $f()$ is the density of a distribution. Non-specific distributions are denoted as $\mathcal{G}$, $\mathcal{F}$ and $\mathcal{E}$. Indices are used to distinguish different distributions behind $\mathcal{G}$ (or $\mathcal{F}$, $\mathcal{E}$) within one model. The abstract formulation of distributions with $\mathcal{G}$, $\mathcal{F}$ and $\mathcal{E}$ aims to guide the focus on the technical ideas of the model. Explicit distributions, which were proposed by the authors are given as well, however, the exact distributions (including distribution family and parameters) are tuning characteristics in the modular framework and can be adapted according to the purpose of the trial. Explicit distributions are denoted according to known conventions from literature e.g. $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with expected value μ and variance σ^2. Truncated distributions are prefixed with the letter T. The variable ϕ is a general tuning parameter and $\mathbf{\Omega}$ denotes a matrix.

3.1.2 Sharing tools

The statistical tools to share information among the baskets reflect whether a frequentist or a Bayesian approach is applied. Hence, the sharing tools are presented grouped by the underlying statistical technique, which is also a proposed metric to categorize the tools. Additionally, the order of presentation is given by the evolution of the tools and how they are

related among each other. The connections between the sharing tools, how the tools evolved from each other, and a categorization of them are graphically presented in Figure 3.2.

Frequentist: Pool all or nothing

The frequentist tools use the *pool all or nothing* approach. At a predefined interim node of the trial, the hitherto observed data is evaluated and, based on these results, a binary sharing decision is taken. The decision is either to pool all baskets into one combined data set and consider them all together, or to investigate each basket independently from the others with individual analyses and decisions for each basket. The tools proposed in literature to make the binary decision are:

(i) *Pass 1st stage* - Chen et al. (2016), Li et al. (2019), Zhou et al. (2019)
The *pass 1st stage* rule is a combination of futility assessment and implicit homogeneity evaluation. Those baskets that pass the first stage without a futility stop are considered to be homogeneous with respect to the treatment response and therefore pooled into one group after the interim analysis. The baskets that do not pass the first stage are stopped and not further investigated. This rule can be considered as the simplest one.

(ii) *(Fisher's) exact test for contingency table* - Cunanan et al. (2017b)
The (Fisher's) exact test is used to investigate the hypothesis of heterogeneous baskets with respect to the treatment responses. If the null hypothesis H_0 : *homogeneous response rates* is rejected, each basket is investigated individually. If not and the null hypothesis is kept, then all baskets are pooled into one group. The critical value for the test statistic (or its significance level) is a tuning parameter and must be defined in the planning phase of the trial.

Bayesian techniques

The Bayesian techniques share information among the baskets with different methods and different underlying concepts. The available techniques can be divided into three groups. The first group is the Bayesian hierarchical model with normal distribution for the transformed response rate θ_i and is denoted as BHM. The simple BHM is the basic version of Bayesian information sharing in basket trials. Also, it is the base for the evolution of sharing techniques

for basket trials. Thall et al. (2003) already presented it for the analysis of phase II trials with multiple subgroups. The underlying assumption is that the treatment effect among the subgroups is exchangeable and correlated. Berry et al. (2013) adapted this model to the developing field of basket trials. The available BHMs are either mean- or variance-driven. The second group of Bayesian techniques shares the information among the baskets without transforming the response rate and directly uses distributions on p_i. The third group evaluates the probabilities of the underlying hypotheses and uses these to share information among the baskets.

(iii) *BHM* - Thall et al. (2003), Berry et al. (2013)

The basic BHM assumes that each basket-individual, logit-transformed response rate comes from the same normal distribution. This means the baskets are assumed to be exchangeable. The hierarchy in the model is introduced by defining the parameters of the normal distribution (μ and σ^2) as random variables itself, with their own distributions $\mathcal{F}$ and $\mathcal{G}$. The parameters which define $\mathcal{F}$ and $\mathcal{G}$ are the hyperparameters and form the second level of the BHM. The hyperparameters are usually set to fixed values, but they can also be defined as random variables, which then creates another hierarchy level. The BHM of Thall et al. (2003) and Berry et al. (2013) is a two-level hierarchical model, where the hyperparameters are fixed. The expected value μ influences the location of the (transformed) response rates and σ^2 determines the amount of sharing among the baskets. The Bayesian distribution of the transformed response rate θ_i is defined in the *BHM* by

$$\begin{aligned} \theta_i &\sim N(\mu, \sigma^2) \\ \mu &\sim \mathcal{G} \qquad\qquad \forall i = 1, ..., I. \\ \sigma^2 &\sim \mathcal{F} \end{aligned}$$

Berry et al. (2013) transform the response rate to $\theta_i := \text{logit}(p_i) - c$, with a constant reference value c. The constant value c can also be incorporated in the expected value μ such that a logit-transformation $\theta_i := \text{logit}(p_i)$ is used. In the original publication

Berry et al. (2013) propose to use a normal distribution with a mean corresponding to an assumed null value for $\mathcal{G}$ and an inverse gamma distribution for $\mathcal{F}$.

Mean-driven BHMs

The mean-driven BHMs expand the basic BHM with additional BHMs or individual distributions. This covers a wider field of locations for possible distributions and for example allows to separate the support scale of θ into a favourable (null) and a non-favourable (alternative) scenario. The scenarios differ by a shifted expected value and each of the individual BHMs are applied with a probability λ, which can be fixed or random as well.

(iv) *ExNex* - Neuenschwander et al. (2016)

The ExNex (exchangeable - nonexchangeable) design adds to the BHM an individual distribution, which represents the scenario when a basket-individual response rate does not fit to the response rates in the other baskets. In that case non-exchangeability between the baskets is assumed. The distribution of θ_i is then given by

$$\theta_i \sim \begin{cases} N(\mu, \sigma^2), & \text{with probability } \lambda_i \\ \quad \mu \sim \mathcal{G} & \\ \quad \sigma^2 \sim \mathcal{F} & \\ N(m_i, v_i), & \text{with probability } (1-\lambda_i) \\ \quad m_i, v_i \text{ fixed} & \end{cases} \qquad \forall i = 1, ..., I.$$

λ_i fixed

The parameters of the individual distribution m_i and v_i are fixed and predefined values. The probabilities λ_i are fixed *a priori*. The λ_i can, but do not have to be different values for each basket. This can be used when it is already assumed *a priori* that a basket i does not conform with the others. Additionally, Neuenschwander et al. (2016) allow to include another hierarchy level by defining λ_i as a random variable with, e.g. Dirichlet distributions. The authors propose to use a normal distribution for $\mathcal{G}$ and a half-normal prior distribution for σ^2.

(v) *PoC* - Jin et al. (2020a)

The next step in the evolution of information sharing is the weighted mixture of two

BHMs. The proof-of-concept (PoC) design by Jin et al. (2020a) is proposing such an approach. The BHMs differ in the expected value around which they are centered. The variance in both BHMs comes from the same distribution. The probability λ is a random variable and describes how likely each of the BHMs is chosen, it is the same for every basket. The distribution of θ_i is then given by

$$\theta_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{with probability } \lambda \\ \quad \mu_1 \sim \mathcal{G}_1 & \\ \quad \sigma_1^2 \sim \mathcal{F} & \\ N(\mu_2, \sigma_2^2), & \text{with probability } (1-\lambda) \\ \quad \mu_2 \sim \mathcal{G}_2 & \\ \quad \sigma_2^2 \sim \mathcal{F} & \end{cases} \qquad \forall i = 1, ..., I.$$
$$\lambda \sim \mathcal{U}[0,1]$$

The PoC design uses an additional hierarchy level for the parameters of $\mathcal{G}_1$ and $\mathcal{G}_2$. For $\mathcal{G}_1$ and $\mathcal{G}_2$, the authors propose a normal distribution with fixed expected value and inverse gamma distributed variance. For $\mathcal{F}$, they propose an inverse gamma distribution.

(vi) *QBHM* - Liu et al. (2017)

Liu et al. (2017) propose a sharing technique for the final analysis that is very similar to the one in PoC. The differences are individual variances for the BHMs and a fixed probability for each BHM. Liu et al. (2017) use a Cochran's Q test to evaluate homogeneity and only apply the mixture of BHMs if homogeneity was not rejected. If homogeneity is rejected, each basket is investigated independently. This reflects a modified test-then-pool approach where Cochran's Q test determines whether pooling is conducted. If the data is pooled, a partial pooling via the mixture of two BHMs is performed. The design of Liu et al. (2017) is called *QBHM* because of the Cochran's Q test.

$$\theta_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{with probability } \lambda \\ \quad \mu_1 \sim \mathcal{G}_1 \\ \quad \sigma_1^2 \sim \mathcal{F}_1 \\ N(\mu_2, \sigma_2^2), & \text{with probability } (1-\lambda) \\ \quad \mu_2 \sim \mathcal{G}_2 \\ \quad \sigma_2^2 \sim \mathcal{F}_2 \end{cases} \qquad \forall i = 1, ..., I.$$

λ fixed

The authors propose a normal distribution for $\mathcal{G}_1$ and $\mathcal{G}_2$, and a gamma distribution for $\mathcal{F}_1$ and $\mathcal{F}_2$.

(vii) *BaCIS* - Chen and Lee (2019)

The Bayesian classification and information sharing design (BaCIS) allocates in a first step each basket to one of two clusters C_1 or C_2. The clustering process is based on two BHMs with different means and the membership in C_1 or C_2 is determined by the higher frequency of allocation to one of the two BHMs in the clustering process. The second step is to apply a BHM exclusively within each cluster, which means to share information only between baskets from the same cluster. Hence, the BaCIS sharing tool is a binary mixture of two BHMs with a beforehand clustering.

$$\theta_i \sim \begin{cases} N(\mu_1, \sigma_1^2) \mid \mathcal{D}_{c_1}, & \text{if basket i located to cluster } C_1 \\ \quad \mu_1 \sim \mathcal{G}_1 \\ \quad \sigma_1^2 \sim \mathcal{F} \\ N(\mu_2, \sigma_2^2) \mid \mathcal{D}_{c_2}, & \text{if basket i located to cluster } C_2 \\ \quad \mu_2 \sim \mathcal{G}_2 \\ \quad \sigma_2^2 \sim \mathcal{F} \end{cases} \qquad \forall i = 1, ..., I.$$

For $\mathcal{G}_1$ and $\mathcal{G}_2$, the authors propose normal distributions with different means and the same variance. For $\mathcal{F}$, they propose a gamma distribution.

(viii) *BLAST* - Chu and Yuan (2018b)

The Bayesian latent subgroup trial (BLAST) extends the two-element mixtures to a mixture of k BHMs. Each of the k BHMs describes a cluster of baskets with similar treatment effect. The membership of basket i in a cluster is latent and determined by the longitudinal trajectory of a biomarker and by the observed binary treatment responses. The biomarker needs to be correlated with the treatment response and serves as proxy for the latter. The number of clusters k is determined by the goodness of fit in the semi-parametric mixed model for the longitudinal trajectory of the biomarker. Chu and Yuan (2018b) argue that clustering of binary responses in small sample sizes is of limited quality and the authors aim to enrich the binary responses by the biomarker to take more informed decisions.

$$\theta_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{with probability } \lambda_{i1} \\ \quad \mu_1 \sim \mathcal{G}_1 & \\ \quad \sigma_1^2 \sim \mathcal{F} & \\ \quad \vdots & \forall i = 1, ..., I. \\ N(\mu_k, \sigma_k^2), & \text{with probability } \lambda_{ik} \\ \quad \mu_k \sim \mathcal{G}_k & \\ \quad \sigma_k^2 \sim \mathcal{F} & \end{cases}$$

$$\lambda_i \sim \text{Dir}(\gamma_1, ..., \gamma_k)$$

The k-dimensional vector of probabilities λ_i has a Dirichlet distribution. The authors propose a normal distribution for all $\mathcal{G}_i$ and an inverse gamma distribution for $\mathcal{F}$.

Variance-driven BHMs

Apart from the location, given by the expected value, the sharing in the BHM can also be guided by the variance. The variance describes the amount of information that is shared. The larger the variance, the less information is shared among the baskets and vice versa. This is because a large variance allows many values around the expected value, while a small variance restricts the values of θ_i closely to the expected value for all baskets i. The extremes are $\sigma^2 = \infty$ for complete independence of the baskets and $\sigma^2 = 0$ for complete sharing among them.

(ix) *calBHM* - Chu and Yuan (2018a)

The calibrated BHM (*calBHM*) is an adaption of the basic BHM where a fixed and calibrated value for the variance is used. The calculated variance value is a monotonically increasing function of the test statistic $\mathcal{T}$ from the χ^2-test and a tuning parameter ϕ. Simulations are used to calibrate the function and the tuning parameter in order to control the type 1 error. While any monotonically increasing function is possible, the authors proposed $g(\mathcal{T}; \phi = (\phi_1, \phi_2)) := exp(\phi_1 + \phi_2 \cdot log(\mathcal{T}))$ with $\phi_2 > 0$ because they observed a robust behaviour for that monotone function in their simulations. The model is thus given by

$$\begin{aligned} \theta_i &\sim N(\mu, \sigma^2) \\ \mu &\sim \mathcal{G} \qquad\qquad \forall i = 1, ..., I. \\ \sigma^2 &= g(T; \phi) \end{aligned}$$

The authors propose a normal distribution for $\mathcal{G}$. Novelli (2021) proposed a slightly modified calibration function $g(\mathcal{T}; \phi = (\phi_1, \phi_2)) = exp(\phi_1 + \phi_2 \cdot \sqrt{\mathcal{T}}) - 1$ in order to improve the overall trial characteristics.

(x) *corBHM* - Jin et al. (2020b)

Jin et al. (2020b) proposed a multivariate normal distribution with a correlation matrix $\mathbf{\Omega}$ to model the basket-individual response distributions. The model is denoted as *corBHM* and each element ω_{ij} of $\mathbf{\Omega}$ is generated with a correlation function of the pairwise distance measures d_{ij} and a random tuning parameter $\phi \sim \mathcal{E}$. The distances are calculated on the posterior distributions of each individual basket and the authors propose the Kullback-Leibler distance, the Hellinger distance, or the Bhattacharyya distance for the pairwise d_{ij}.

$$\begin{aligned} \theta &\sim MVN(\mu, \mathbf{\Omega} \cdot \sigma_b^2 + \text{diag}(\sigma_w^2)) \\ \mu &\sim \mathcal{G} \\ \sigma_b^2 &\sim \mathcal{F}_b, \quad \sigma_w^2 \sim \mathcal{F}_w \end{aligned}$$

The variable σ_b^2 is the variance between the baskets. The variance within a basket σ_w^2 is stored in the diagonal matrix $\text{diag}(\sigma_w^2)$. *corBHM* applies an additional hierarchy level for the variance of $\mathcal{G}$. The authors propose a normal distribution for $\mathcal{G}$ with a fixed expected value and a variance with inverse gamma distribution. For $\mathcal{F}_b$ and $\mathcal{F}_w$, they propose inverse gamma distributions.

(xi) *BCHM* - Chen and Lee (2020)

The sharing of information in the *BCHM* design consists of two steps. The first step is to cluster the baskets with a Dirichlet process mixture model (DPM). For each basket, the clustering process returns the relative frequency of a pairwise membership in the same cluster during the sampling process. The relative frequency is denoted by ω_{ij} and describes the similarity of two baskets with respect to the response rate. In the second step, BHMs with respect to each basket i are constructed and the reference to i is denoted as a superscripted index in round brackets. The ω_{ij} modify the variance in the BHMs and consequently determine the degree of sharing between the baskets. This leads to an individual variance for basket j in the BHM with respect to basket i. The variance modification is the quotient of the variance of basket i, ${\sigma^{(i)}}^2$, and the pairwise similarity measure ω_{ij} to basket j. Hence, the higher the similarity measure ω_{ij}, the smaller the variance and the higher the pairwise sharing. The distribution of θ_i is then given by the i-th element from the BHM with respect to i and is denoted by

$$\begin{aligned} \theta_j^{(i)} &\sim N\Big(\mu^{(i)}, \tfrac{{\sigma^{(i)}}^2}{\omega_{ij}}\Big) \qquad \forall j = 1, ..., I \\ \mu^{(i)} &\sim \mathcal{G} \\ {\sigma^{(i)}}^2 &\sim \mathcal{F} \\ \theta_i &:= \theta_i^{(i)} \qquad \forall i = 1, ..., I. \end{aligned}$$

The authors propose a normal distribution for $\mathcal{G}$ and an inverse gamma distribution for $\mathcal{F}$.

(xii) *Zheng* - Zheng and Wason (2020)

In the design of Zheng and Wason (2020), a marginal predictive prior (MPP) is used to share information among the baskets. The MPP is a weighted linear combination

of $(I-1)$ random variables and created for each basket i. The elements in the linear combination are commensurate predictive priors (CPP_{ij}) for basket i based on basket j, $j \neq i$ and the weights ω_{ij} are functions of the Hellinger distance between basket i and j. A CPP_{ij} element is created by integrating over the variance σ_{ij}^2 of a normally distributed random variable θ_{ij}. The expected value μ_j of this normal distribution is fixed and inferred from the observed results in basket j. The variance σ_{ij}^2 is a random variable and its distribution depends on the similarity between the results in basket i and j and therefore controls the information sharing from j to i. In the analysis of basket i, the MPP serves as the prior and is updated by the observed data in basket i.

$$\theta_{ij} \sim N(\mu_j, \sigma_{ij}^2)$$

$$\mu_j \text{ fixed and inferred from results in basket } j$$

$$\sigma_{ij}^2 \sim \mathcal{F}(d_{ij})$$

CPP_{ij}: integrate over the distribution of the variance σ_{ij}^2 which results in

$$\theta_j^{(i)} \sim N(m_j, \sigma_j^2)$$

with inferred and non-random values for μ_j, σ_j^2

MPP: combine all $(I-1)$ CPPs in a linear combination of random variables

$$\theta_i := \textstyle\sum_{j \neq i} \omega_{ij} \cdot \theta_j^{(i)} \sim N(\sum_{j \neq i} \omega_{ij} \cdot m_j, \sum_{j \neq i} \omega_{ij}^2 \cdot \sigma_j^2) \qquad \forall i = 1, ..., I.$$

For the distribution of σ_{ij}^2, the authors propose the quadratic inverse of a spike-and-slab distribution (Mitchell and Beauchamp, 1988) based on the Hellinger distance d_{ij}. The spike-and-slab distribution in this model is a mixture of a point mass and a truncated uniform distribution. The point mass lies above the support scale of the truncated uniform distribution and induces strong sharing whereas the uniform distribution covers situations where sharing is not indicated.

Non-transformed and beta-binomial model

The so far presented BHMs are all based on the transformed response rate θ_i, however, there are sharing techniques which directly work with the non-transformed response rate p_i. The non-transformed techniques use the conjugate combination of binomial data and beta distributed response rates which result in a beta-binomial model.

(xiii) *Simon* - Simon et al. (2016)

Simon's basket trial works with a non-transformed and categorical response rate $p \in \{p_0, p_1\}$ and the basket-individual p_i takes the value p_0 with probability γ. The notation for the categorical distribution is $Cat(p_0, p_1; \gamma)$. The information sharing by Simon consists of two scenarios. In the homogeneous scenario, exchangeability of the response rates among all baskets is assumed. Hence, all baskets are evaluated together. In the heterogeneous scenario, independence between the baskets is assumed and each basket is evaluated individually. The two scenarios and their assumptions are similar to the ExNex approach, however, Simon uses categorical response rates and the data are fully shared while ExNex uses a continuous scale and allows an adjusted degree of sharing in the BHM. The separation into two scenarios (complete sharing or individual evaluation) is a Bayesian *all or nothing* approach.

$$p_i \sim \begin{cases} Cat(p_0, p_1; \gamma) \mid \mathcal{D}, & \text{with probability } \lambda \\ Cat(p_0, p_1; \gamma) \mid \mathcal{D}_i, & \text{with probability } (1-\lambda) \end{cases} \qquad \forall i = 1, ..., I.$$

λ, γ fixed

Parameter λ is a fixed value and the same for all baskets, without any further specifications in the original paper. The posterior probability of p_i can be calculated in closed form using the Bayes theorem.

(xiv) *Asano* - Asano and Hirakawa (2020)

The sharing tool of Asano and Hirakawa (2020) is an extension of the Bayesian *all or nothing* approach by Simon et al. (2016). However, Asano and Hirakawa (2020) use a continuous scale on $(0, 1)$ for the response rate p, and in the heterogeneous scenario, they use two different priors for the individual analyses of the baskets. The two different priors reflect different prior knowledge on the response rate and the first prior is chosen with fixed probability ω. The fixed parameter λ describes the probability for the homogeneous scenario. The authors originally presented their design in form of a Bayesian model averaging approach as proposed by Psioda et al. (2019). The distribution of the basket individual response rates are then given by

$$p_i \sim \begin{cases} Beta\Big(a + \sum_j r_j, b + \sum_j (n_j - r_j)\Big), & \text{with probability } \lambda \\ Beta\big(a_1 + r_i, b_1 + (n_i - r_i)\big), & \text{with probability } (1-\lambda)\cdot\omega \\ Beta\big(a_2 + r_i, b_2 + (n_i - r_i)\big), & \text{with probability } (1-\lambda)\cdot(1-\omega) \end{cases} \qquad \forall i = 1, ..., I.$$

λ, ω fixed

(xv) *MEM* - Hobbs and Landin (2018)

The multi-exchangeability model (*MEM*) shares information between the baskets based on an exchangeability matrix $\mathbf{\Omega}$. The matrix $\mathbf{\Omega}$ consists of binary entries, each displaying in a pairwise manner whether two baskets are exchangeable ($\omega_{ij} = 1$) or not ($\omega_{ij} = 0$). The exchangeability matrix $\mathbf{\Omega}$ is random and depends on the distribution of the matrix elements ω_{ij}. The authors use *a priori* equal probabilities $\mathrm{P}[\omega_{ij} = 1] = 0.5$ for all pairwise entries in their example. The information of two baskets is combined if $\mathbf{\Omega}$ indicates exchangeability between the two baskets, while elsewise they remain separated. The MEM design iterates over all possible exchangeable-nonexchangeable combinations for every basket and models them together with the observed data. The resulting beta-binomial model describes the posterior distributions of the basket-wise response rates

$$p_i \sim Beta\Big(a + \sum_j \omega_{ij} \cdot r_j, b + \sum_j \omega_{ij} \cdot (n_j - r_j)\Big) \qquad \forall i = 1, ..., I.$$

a, b fixed values

$\mathbf{\Omega} \sim \mathcal{E}$

(xvi) *BMA* - Psioda et al. (2019)

The Bayesian model averaging (BMA) of Psioda et al. (2019) presents the same idea as the *MEM* approach by Hobbs and Landin (2018). Every possible combination of pairwise exchangeability among the baskets is modeled together with the observed data. The posterior distributions are then calculated with a beta-binomial model.

(xvii) *Fujikawa* - Fujikawa et al. (2020)

The Fujikawa design shares information via the properties of the beta-binomial model. The knowledge and observations from the other baskets are incorporated in weighted form to the posterior distribution of p_i. The pairwise weights ω_{ij} are a function of the Jensen-Shannon divergence of the individual posterior distributions from basket i and j, and of two fixed tuning parameters. The basket-individual beta priors are given by the fixed values a_j and b_j. The distribution of the basket-individual response rate is then described by

$$p_i \sim Beta\Big(\sum_j \omega_{ij} \cdot (a_j + r_j), \sum_j \omega_{ij} \cdot (b_j + n_j - r_j)\Big) \qquad \forall i = 1, ..., I.$$

a_j, b_j fixed values

Hypothesis-driven models

The hypothesis-driven sharing tools use the probabilities for the null and alternative hypothesis and do not directly rely on the response rate. Information is shared among those baskets with similar probabilities for the respective hypotheses.

(xviii) *RoBoT* - Zhou and Ji (2020)

The robust Bayesian hypothesis testing method (*RoBoT*) design shares information within latent subgroups. A Dirichlet process mixture model (DPM) determines the latent subgroups based on the probabilities of basket-individual hypotheses ($H_0 : p_i \leq p_{i0}$, $H_1 : p_i > p_{i0}$). The defining parameters of the Dirichlet process $\mathcal{DP}$ are the fixed concentration parameter α and the base distribution G_0. Hence, the distribution of the basket-indvidual response rate p_i is given by

$$p_i \sim \begin{cases} TBeta(a_0, b_0, [0, p_{i0}]), & \text{with probability } \lambda_i \\ \quad a_0, b_0 \text{ fixed} & \\ TBeta(a_1, b_1, (p_{i0}, 1]), & \text{with probability } (1-\lambda_i) \\ \quad a_1, b_1 \text{ fixed} & \end{cases} \qquad \forall i = 1, ..., I.$$

$$\text{logit}(\lambda_i) \sim N(\mu_i, \sigma_i^2)$$
$$(\mu_i, \sigma_i^2) \sim G$$
$$G \sim \mathcal{DP}(\alpha, G_0)$$

The authors propose a two-dimensional distribution as a combination of a normal and truncated Cauchy distribution for G_0.

(xix) *MUCE* - Lyu et al. (2020)

Information sharing in the *MUCE* design is based on a hierarchical model for the variable Z_i. This variable is used to indicate the basket-individual hypotheses. The null ($H_0 : \theta_i \leq \theta_{i0}$) and alternative ($H_1 : \theta_i > \theta_{i0}$) hypothesis are chosen based on the distribution of Z_i which means the hierarchical model shares information about the certainty of the hypotheses among the baskets.

$$\theta_i \sim \begin{cases} TCauchy(\theta_{i0}, \gamma, (-\infty, \theta_{i0}]), & \text{with probability } \mathrm{P}[Z_i < 0] \\ \quad \gamma \text{ fixed} & \\ TCauchy(\theta_{i0}, \gamma, (\theta_{i0}, \infty]), & \text{with probability } \mathrm{P}[Z_i \geq 0] \\ \quad \gamma \text{ fixed} & \end{cases} \qquad \forall i = 1, ..., I.$$

$$Z_i \sim N(\mu_i, \sigma_Z^2), \qquad \sigma_Z^2 \text{ fixed}$$
$$\mu_i \sim N(\mu_0, \sigma_0^2), \qquad \sigma_0^2 \text{ fixed}$$
$$\mu_0 \sim N(\mu, \sigma^2) \qquad \mu, \sigma^2 \text{ fixed}$$

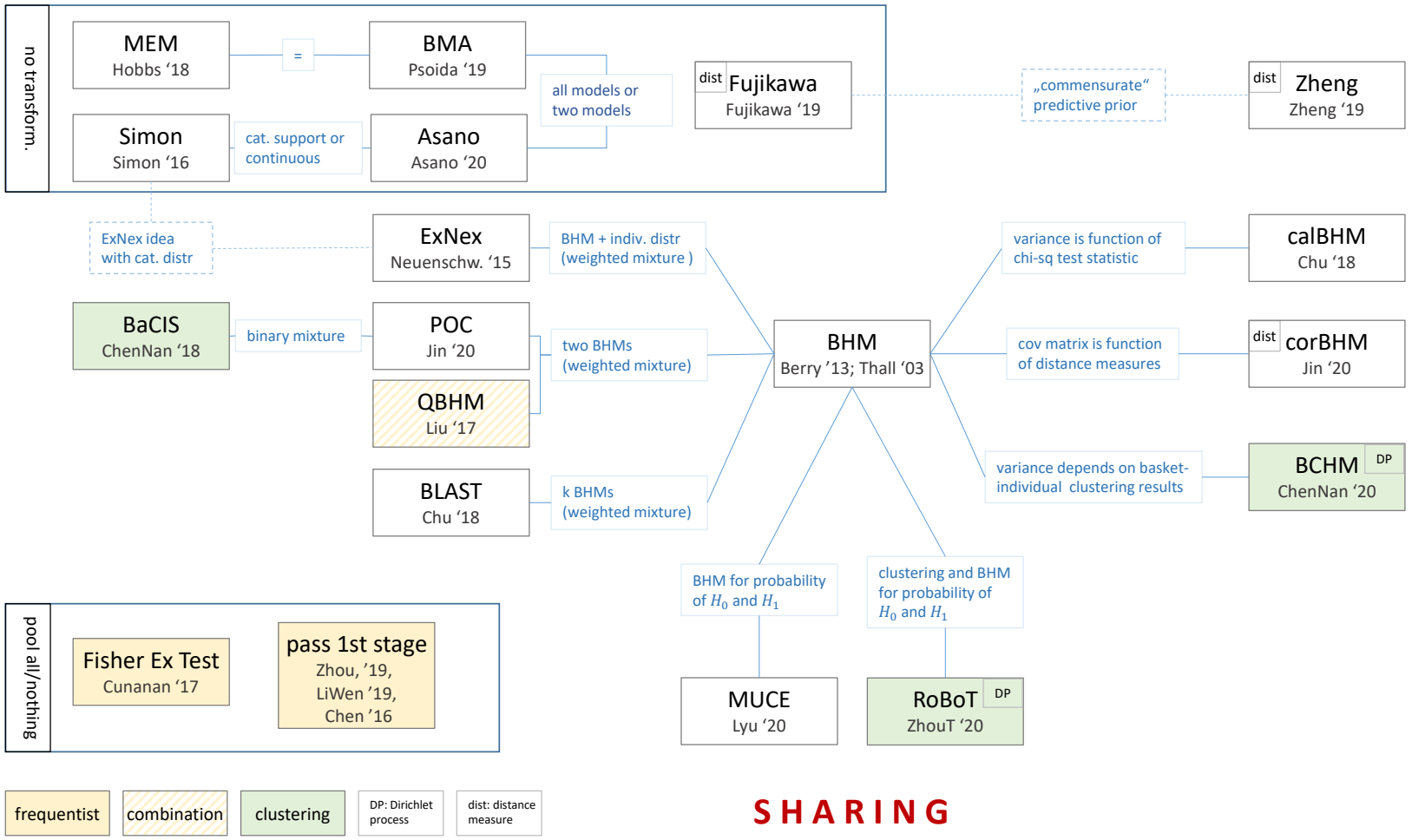


Figure 3.2: *A sharing tool is denoted by a box with name, author, and year of publication. No filling describes a Bayesian technique, yellow filling stands for a frequentist technique and green indicates that a clustering is part of the sharing technique. Connections between techniques are displayed by lines and the modifications are described within the small boxes. Adapted from Pohl et al. (2021).*

3.1.3 Futility tools at interim

In the following, tools for the second component, the interim futility assessment, are presented. The tools are defined rules to make an informed decision whether to stop the recruitment to the basket at the interim node or not. The tools all work with the non-transformed response rate p_i. Hence, designs which use the transformed response rate to model the posterior distributions re-transform back to p_i to take the interim futility decision. Detailed technical relationships between the interim futility tools are elaborated in Section 3.2.3.

(i) **Minimum number of responses:** Prune basket i if less than r_i responses have been observed.
Used by: Cunanan et al. (2017b), Zhou et al. (2019)

(ii) **Statistical test:** Prune basket i if the p-value from an appropriate statistical test for the hypotheses of the primary endpoint (e.g. $H_0 : p \leq p_0$, $H_1 : p > p_0$) exceeds α_F. The significance level is a tuning parameter of the basket trial design and predefined. The value for α_F should be higher than the commonly used 5% for statistical tests, because otherwise the pruning would be too strong and many baskets would stop early.
Used by: Chen et al. (2016)

(iii) **Posterior probability:** Prune basket i if the posterior probability that the response rate exceeds the fixed reference value p_0 is low, i.e.

$$\mathrm{P}[p_i > p_0 | \mathcal{D}] \leq \tau_F. \tag{3.1}$$

Used by: Hobbs and Landin (2018), Thall et al. (2003), Asano and Hirakawa (2020)
One modification is shift the reference value using two response values (e.g. null and alternative values). The fraction $\frac{p_0+p_1}{2}$ includes both and assesses in which direction the basket-individual response rate tends to, that is

$$\mathrm{P}\left[p_i > \frac{p_0 + p_1}{2} \middle| \mathcal{D}\right] \leq \tau_F. \tag{3.2}$$

Used by: Berry et al. (2013), Chu and Yuan (2018a), Chu and Yuan (2018b), Psioda et al. (2019), Jin et al. (2020b)

The second modification for the posterior probability assumes a discrete distribution of the response rate with only two values p_0 or p_1, i.e.

$$\mathrm{P}[p_i = p_1|\mathcal{D}] \leq \tau_F. \tag{3.3}$$

Used by: Simon et al. (2016)

(iv) **Posterior probability of hypotheses:** Prune basket i if the posterior probability that the response rate follows a distribution according to the alternative hypothesis is low. In the MUCE design of Lyu et al. (2020), the random variable Z_i indicates the distribution of the response rate in basket i. For $Z_i \geq 0$, the basket has a distribution given by the alternative hypothesis, while elsewise the distribution is given according to the null hypothesis. Basket i is then pruned if

$$\mathrm{P}[Z_i \geq 0|D] < \tau_F.$$

Used by: Lyu et al. (2020)

(v) **Posterior predictive probability:** Prune basket i if the posterior predictive probability that the response rate exceeds the reference value at the final analysis is low. It requires that the final number of observations per basket is already known at the interim futility assessment. Liu et al. (2017) simulate the posterior predictive probability. At the interim assessment they use all the available observations and build a beta-binomial model from which they draw numbers of stage two responses for the simulation process. The simulated number of responses are then used to calculate final analyses. The frequency of simulated outcomes for which the response rate in the final analysis is larger than the reference response rate p_0 is the simulated posterior predictive probability. A basket is then pruned if the simulated posterior predictive probability is lower than a predefined threshold τ_F. Fujikawa et al. (2020) applies a beta-binomial model as well, but they compute the posterior predictive probability analytically. A basket is then pruned if the following holds true

$$\sum_{\{r_i:\text{final analysis successful}\}} \int_0^1 f(r_i|p_i) f(p_i|\mathcal{D}) dp_i \leq \tau_F. \tag{3.4}$$

Used by: Liu et al. (2017), Fujikawa et al. (2020)

(vi) **Conditional power:** Prune basket i if the conditional power of a successful final basket analysis is low. It requires that the final number of observations per basket is already known at the interim futility assessment and an assumption about the true response rate must be made. A basket is then pruned if the following holds true

$$\sum_{\{r_i:\text{final analysis successful}\}} f(r_i|\hat{p}_i) \leq \tau_F. \tag{3.5}$$

The assumed response rate $\hat{p}$ at the interim assessment can take any fixed value on the support of p. Informed choices for the point estimate of the assumed response rate can be the maximum likelihood estimate (MLE) based on the so far observed data at interim (Saville et al., 2014), or the initially assumed response rate under the alternative hypothesis. Also, the boundaries of (e.g. 90%) confidence intervals can be informative while still at least partly incorporating the uncertainty regarding the true response rate.
Used by: so far not applied in proposed basket designs but a suitable frequentist tool

In the publications of Zheng and Wason (2020), Li et al. (2019), and Chen and Lee (2019) the interim futility assessment is optional and no further explicit explanations are given. These works concentrate on the method to share information and not explicitly on the complete design. In Neuenschwander et al. (2016), Jin et al. (2020a), Chen and Lee (2020), Zhou and Ji (2020) interim futility assessments are not mentioned but in a modular basket trial their sharing tools can be applied in basket trial designs where futility is investigated at interim.

The frequency of interim futility assessments is a design element and is an optional component. If a futility analysis is incorporated, the time points and the rules must be specified in advance. Jin et al. (2020b), Liu et al. (2017), Chen et al. (2016), Cunanan et al. (2017a), and Zhou et al. (2019) propose one interim futility assessment whereas Simon et al. (2016), Chu and Yuan (2018b), Chu and Yuan (2018a), Psioda et al. (2019), Hobbs and Landin (2018), Fujikawa et al. (2020), Li et al. (2019), and Lyu et al. (2020) allow multiple assessments. The time points for the assessments can depend on the passed trial time or on the amount of observed data. In the design of Berry et al. (2013), the futility assessments take place after a certain number (e.g. 10) of patients per basket have been observed. Additional interim looks are

then planned after five additional patients were observed. The most extreme scenario is an interim futility assessment after each observation (Simon et al., 2016) while the most liberal scenario only conducts one single interim futility assessment (e.g Jin et al. (2020b)) which reflects a two-stage design.

3.1.4 Efficacy tools at interim

The so far proposed tools for the third component are presented in the following. The tools describe rules to stop baskets at an interim assessment due to strong evidence of a successful treatment. The proposed tools for the interim efficacy assessments are:

(i) **Statistical test:** Prune basket i if the p-value from an appropriate statistical test for the primary endpoint exceeds α_E. The significance level is a tuning parameter of the basket trial design and predefined in accordance with the regulatory requirements for a potential accelerated approval of this basket.
Used by: Chen et al. (2016)

(ii) **Posterior probability:** Prune basket i if the posterior probability that the response rate exceeds the reference value is high, i.e.

$$\mathrm{P}[p_i > p_0|\mathcal{D}] > \tau_E. \tag{3.6}$$

Used by: Psioda et al. (2019)

Equivalently to the futility assessments, the reference value can be changed to

$$P\left[p_i > \frac{p_0 + p_1}{2}\middle|\mathcal{D}\right] > \tau_E. \tag{3.7}$$

Used by: Berry et al. (2013)

The posterior probability rule can be adapted for a discrete distribution of the response rate, resulting in

$$\mathrm{P}[p_i = p_1|\mathcal{D}] > \tau_E. \tag{3.8}$$

Used by: Simon et al. (2016)

(iii) **Posterior predictive probability:** Prune basket i if the posterior predictive probability that the response rate exceeds the reference value at the final analysis is high, i.e.

$$\sum_{\{r_i:\text{final analysis successful}\}} \int_0^1 f(r_i|p_i) f(p_i|\mathcal{D}) dp_i > \tau_E. \tag{3.9}$$

Used by: Fujikawa et al. (2020)

The interim efficacy assessment is optional and is explicitly proposed only in a few designs (Chen et al., 2016; Simon et al., 2016; Psioda et al., 2019; Fujikawa et al., 2020). In Berry et al. (2013), Chen and Lee (2019), and Zheng and Wason (2020) it is mentioned to be incorporated into the design, but without any further technical definitions in the former two. The other designs do not consider interim efficacy assessments.

The tools are very similar to the futility tools. The difference is in the direction from which the thresholds are considered. The operand $<$ is changed to $>$ for the efficacy tools (cf. Equation 3.2 and 3.7, Equation 3.1 and 3.6, Equation 3.3 and 3.8, and Equation 3.4 and 3.9). In the BMA design by Psioda et al. (2019), different reference values in the futility and efficacy assessments are applied. In the other designs, the reference value remains consistent and only the direction of the threshold is changed. The frequency of interim efficacy assessments is a design element and the time points for these assessments must be prespecified. A pragmatic approach is to conduct them simultaneously with the interim futility assessment. Still, the efficacy assessment can also take place independently at nodes where it is a meaningful tool for the purpose of the trial. Early trials rather concentrate on pruning baskets with no effect, in order to save resources and to only continue with potentially promising indications. In that case, the interim efficacy assessment is of limited benefit but leaves the door open for a quick development of a potential breakthrough treatment with overwhelming early results.

3.1.5 Final analysis

The so far proposed tools for the fourth component are presented in the following. The tools evaluate in the final analysis the success of the basket trial. The final analysis focuses on the underlying research question (confirmatory evaluation of the investigated treatment, detection of promising baskets, PoC) and accordingly uses the observed results in each basket and throughout the complete trial.

(i) **Statistical test:** The frequentist designs apply a statistical test to assess the final efficacy. A binomial test can be applied for the analysis of binary response rates. Cunanan et al. (2017b) apply the one-sided version ($H_0 : p \leq p_0, H_1 : p > p_0$) of the test and Zhou et al. (2019) use the two-sided version ($H_0 : p = p_0, H_1 : p \neq p_0$). The more general approaches for the confirmatory phase III basket trials (Chen et al., 2016; Li et al., 2017) only refer to an appropriate statistical test for the investigated primary endpoint.

(ii) **Posterior probability:** The final efficacy analysis is conducted on the posterior distribution of the response rate. The probability that the response rate in basket i exceeds a reference value is compared with the threshold for the final analysis τ_A. Basket i is considered promising if it fulfills

$$\mathrm{P}[p_i > p_0|\mathcal{D}] > \tau_A. \tag{3.10}$$

Used by: Berry et al. (2013), Neuenschwander et al. (2016), Liu et al. (2017), Hobbs and Landin (2018), Chu and Yuan (2018a), Psioda et al. (2019), Chen and Lee (2019), Chu and Yuan (2018b), Jin et al. (2020b), Asano and Hirakawa (2020)

Similar to the interim futility and efficacy tools, modifications of the posterior probability decision rule are feasible, like for example, equality of the posterior distribution with the threshold τ_A to declare efficacy in the final analysis

$$\mathrm{P}[p_i > p_0|\mathcal{D}] \geq \tau_A. \tag{3.11}$$

Used by: Fujikawa et al. (2020)

Another modification is to change the reference value and increase the reference value p_0 by the positive value δ to incorporate an additional margin which must be exceeded to declare a basket successful, i.e.

$$\mathrm{P}[p_i > p_0 + \delta|\mathcal{D}] > \tau_A.$$

Used by: Zheng and Wason (2020), Chen and Lee (2020)

In the case of a discrete distribution with only two values p_0, p_1 for the response rate, the posterior probability can be adapted to evaluate the probability of the alternative value p_1, i.e.

$$\mathrm{P}[p_i = p_1|\mathcal{D}] > \tau_A.$$

Used by: Simon et al. (2016)

In the hypothesis-driven design of Lyu et al. (2020), the random variable Z_i is investigated because this variable uniquely indicates whether the response rate takes values larger or smaller than the reference value p_0. Hence, in the final analysis the posterior probability of the alternative hypothesis is evaluated which is equal to Equation 3.10

$$\mathrm{P}[Z_i \geq 0|\mathcal{D}] > \tau_A. \tag{3.12}$$

Used by: Lyu et al. (2020)

The information $\mathcal{D}$, which is used in the final analysis, depends on the observations, but also on the sharing tools applied throughout the trial. Therefore, $\mathcal{D}$ contains a different amount of information among different basket trial designs. For example in Chen and Lee (2019), only the information from the cluster to which the basket belongs to is considered, while others, like Fujikawa et al. (2020), use all the available information.

(iii) **Proof-of-concept probability:** The PoC design evaluates all baskets together in the final analysis with the goal to state that at least one basket shows efficacious results. A single basket fulfills this requirement if its posterior response rate is distributed according to the alternative hypothesis. The alternative response rate distribution is represented by one of the two BHMs (cf. sharing of PoC design). Hence, the final analysis is conducted with

$$\mathrm{P}[\text{at least one efficacious basket}|\mathcal{D}] = P\left[\left(\sum_i^I \mathbb{1}_{\{\text{basket } i \text{ efficacious}\}}\right) > 0 \middle| \mathcal{D}\right] > \tau_A.$$

Used by: Jin et al. (2020a)

Zhou and Ji (2020) do not propose an explicit tool for the final analysis in their RoBoT design. Their focus rather lies on the sharing tool, but still a final analysis can be conducted with an appropriate tool, for example a tool similar to Equation 3.12, because the RoBoT design is also hypothesis-driven.

3.2 Relationships among decision tools in basket trials

The previous section showed that Basket trials apply frequentist and Bayesian methodology to evaluate the investigated treatment. In this section, connections between frequentist and Bayesian decision rules are investigated. The decisions with respect to futility and/or efficacy of a basket are made at interim and also in the final analysis. This section consists of three parts. The first elaborates the analytical relationship between the binomial test and the posterior probability of the beta-binomial model, and quantifies their discrepancy especially for a uniform prior. With this knowledge, congruence of decisions can be achieved which is applied in the second part by converting the frequentist futility and efficacy decisions in the design of Cunanan et al. (2017b) into Bayesian ones. The third part extends the map of connections among decision tools by the remaining techniques which were introduced in the Section 3.1.

3.2.1 Analytical relationship of one-sided binomial test and Bayesian beta-binomial model

The assessment of futility or efficacy can be made by a frequentist one-sided binomial test or in a Bayesian approach with a decision based on the posterior distribution (cf. Section 3.1). The node of the trial where the decision is taken can be both at interim or at final assessment. The motivation to investigate the connection between frequentist and Bayesian decision tools lies in the type 1 error control of the frequentist technique and the requirement to control the rate of false positive decisions of a Bayesian tool. Moreover, a thorough understanding of both techniques, their characteristics, and the interaction between them is needed to correctly apply the tools in the setting of basket trials.

The formal initial situation is the one-sided binomial test with respect to a significance level $\alpha = 0.05$. The null and alternative hypothesis for the response rate p are $H_0 : p \leq p_0$ and $H_1 : p > p_0$ with regard to the fixed reference value p_0. In the Bayesian setting, the response rate p follows a beta distribution with parameters a and b. The combination of data from a binomial distribution and a beta prior distribution for the response rate results in the conjugate beta-binomial model (Section 2.2.2) with posterior distribution $Beta(a+r, b+n-r)$.

The frequentist decision relies on the p-value and rejects the null hypothesis if the p-value is below the prespecified significance level α. Consequently, for the one-sided binomial test, the p-value has the following form

$$\mathrm{P}[R \geq r|n, p_0] = \sum_{x=r}^{n} \mathrm{P}[R = x|n, p_0] = \sum_{x=r}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x} \tag{3.13}$$

and the alternative hypothesis is accepted if

$$\mathrm{P}[R \geq r|n, p_0] \leq \alpha \tag{3.14}$$

is fulfilled. The Bayesian decision is made on the posterior distribution of the response rate, based on which the probability of the alternative hypothesis is calculated. This probability is compared with the threshold τ. Consequently, the posterior probability to make a decision is

$$\mathrm{P}[p > p_0|n, r] = \frac{1}{B(a+r, b+n-r)} \int_{p_0}^{1} p^{a+r-1}(1-p)^{b+n-r-1} dp \tag{3.15}$$

and is taken in favor of a response rate exceeding the reference value p_0 if

$$\mathrm{P}[p > p_0|n, r] \geq \tau \tag{3.16}$$

holds true (cf. Section 3.1 and Fujikawa et al. (2020)).

The first step to analyse the analytical relationship between the one-sided binomial test and the Bayesian decision is to investigate whether there is a global choice of τ such that the probability of false positive decisions with the Bayesian decision rule is equivalent to the type 1 error of the frequentist test. The T1E is calculated by the sum over all likelihoods of data which result in a decision in favour of H_1 under the assumption that H_0 holds

$$\sum_{r=0}^{n} \mathrm{P}[R = r|n, p_0] \cdot \mathbb{1}_{\{\text{decision for } H_1\}}. \tag{3.17}$$

The following counter-example shows that there is no global τ to control the T1E simultaneously. For the assumed scenario $p_0 = 0.15$ and $\tau = 0.975$ with uniform prior ($a = b = 1$) the Bayesian T1E

- is above 0.05 for some $n < 20$
- is controlled for $n \in [20, 42]$
- and smaller than the T1E of the frequentist decision if $n > 42$.

The evolution of the T1Es in relation to n for this counterexample is shown in Figure 3.3a. A change in the reference value p_0 results in different type 1 errors (see Figure 3.3b) and different regions of n in which the T1E is controlled by the fixed value threshold $\tau = 0.975$.

The figures graphically display the dependence of threshold τ on the number of observations n and on the reference value p_0. This dependency can also be read from equations 3.15 and 3.17 as the posterior distribution is described by n, r, and p_0, and the T1E is a sum over each possible r. Hence, a global control with τ is not possible, however, in local regions of n and for fixed p_0 a local control can be achieved. The threshold must generally be considered as a function of n and p_0 to control the T1E with the Bayesian decision tool.

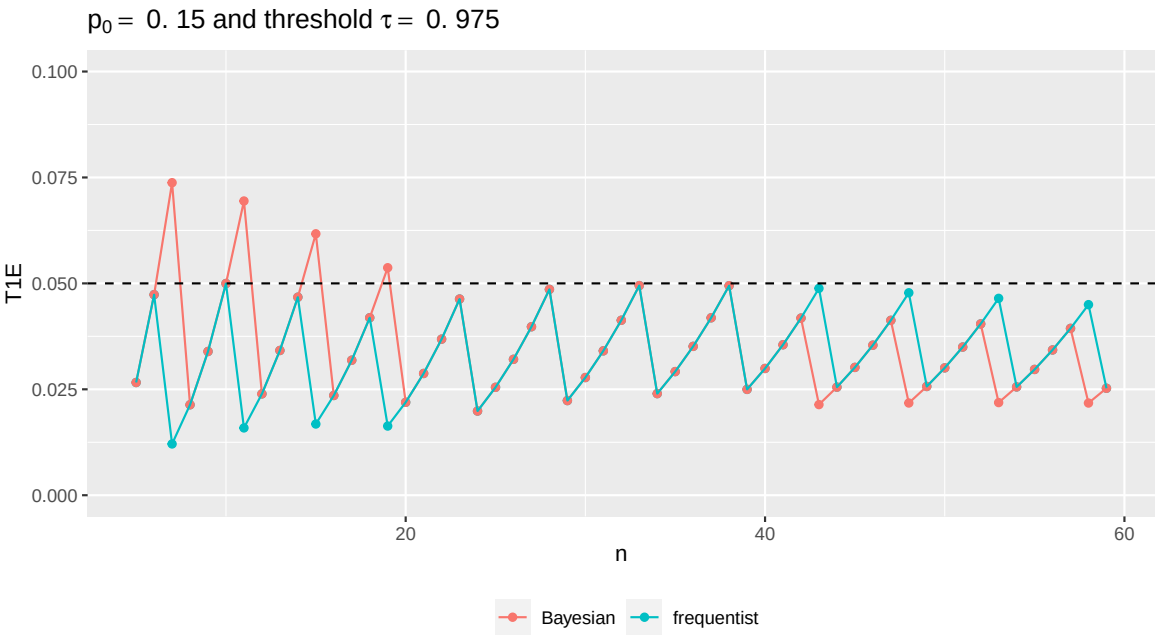
Consequently, the next step is to determine τ such that the Bayesian tool comes to the same decision as the frequentist test. The lowest number of responses r_{min} for which the null hypothesis is rejected with the binomial test is defined as

$$r_{min} := min\{r|\mathrm{P}[R \geq r|n, p_0] \leq \alpha\}$$

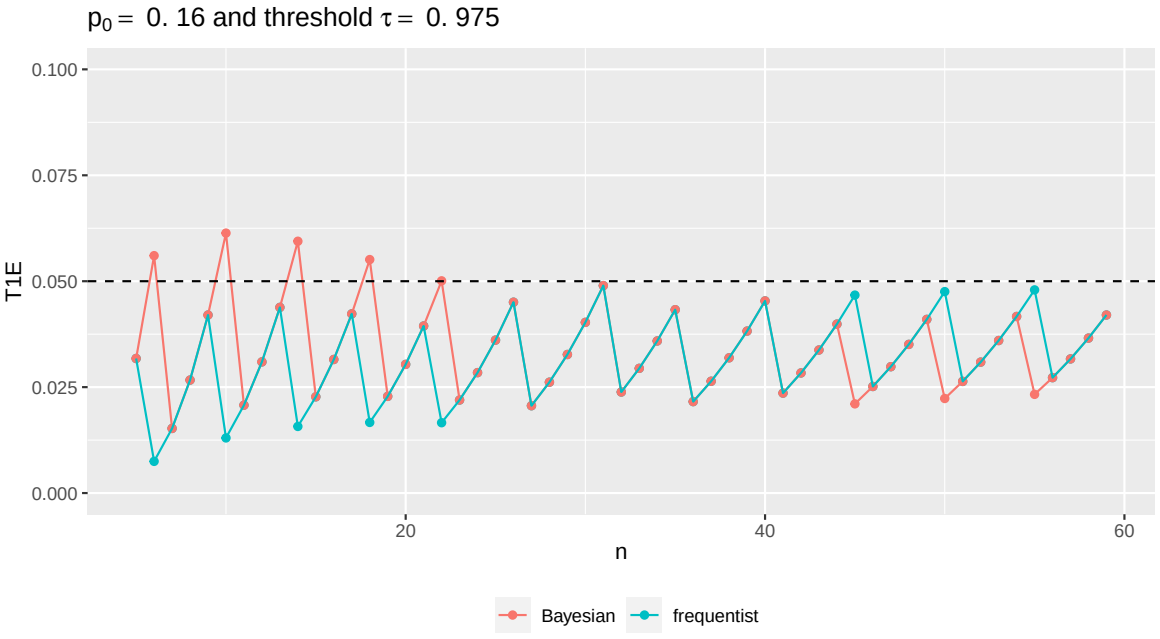
For all $r \geq r_{min}$ it holds that the p-value is less than or equal to the significance level α, hence, it suffices to concentrate on r_{min} as the relevant value to determine an appropriate τ. This monotonicity in r also holds for the Bayesian decision on the posterior probabilities as $\mathrm{P}[p > p_0|n, r+1] \geq \mathrm{P}[p > p_0|n, r] \quad \forall r \in \{0, n-1\}$, for which a proof can be found in Kopp-Schneider et al. (2019). Consequently one has to choose τ such that $\mathrm{P}[p > p_0|n, r] \geq \tau \quad \forall r \geq r_{min}$. This condition is fulfilled for

$$\begin{aligned} \tau_{(n,p_0)} :&= \mathrm{P}[p > p_0|n, r_{min}] = \\ &= c \cdot \int_{p_0}^{1} p^{a+r_{min}-1}(1-p)^{b+n-r_{min}-1} dp. \end{aligned} \tag{3.18}$$

Figure 3.3: *Type 1 error of a one-sided frequentist binomial test with* $\alpha = 0.05$ *and Bayesian decision from a beta-binomial model with uniform prior* $a = b = 1$ *and threshold* τ *with different reference response rates* p_0.



(a) *reference response rate* $p_0 = 0.15$



(b) *reference response rate* $p_0 = 0.16$

with c being the normalizing constant. Finally with τ being chosen in dependence of the total number of observations n and the reference value p_0, the Bayesian tool comes to the same decision as the frequentist test. Therefore, the Bayesian false-positive rate is controlled as in the frequentist binomial test because the indicator function in Equation 3.17 is 1 if $r \geq r_{min}$. In such a case the p-value is smaller than α and at the same time the posterior probability is greater than or equal to $\tau_{(n,p_0)}$.

The latter has shown that congruence between the frequentist test decision and the Bayesian posterior probability decision can be achieved if the threshold τ (or significance level α) is chosen appropriately. It is of further interest whether the existing connection between the two decision tools can be quantified analytically. Therefore, the p-value of the binomial test and the posterior probability are investigated in more detail. The p-value of the binomial test has the appealing characteristic that it can be converted to a function of similar structure as the density of a beta distribution. This profits from the known connection of the binomial cumulative density function and an incomplete beta distribution (e.g. Hartley and Fitch, 1951 or Lieberman and Owen, 1961, p.18). The conversion of the p-value is achieved by deriving the p-value (Equation 3.13) with respect to p_0 using the product rule for derivation and the calculation rules for the binomial coefficient which results in

$$\begin{aligned}\frac{d}{dp_0}\mathrm{P}[R \geq r|n,p_0] &= \sum_{x=r}^{n}\left[\binom{n}{x}x \cdot p_0^{x-1}(1-p_0)^{n-x} - (n-x)\cdot p_0^{x}(1-p_0)^{n-x-1}\right] = \\ &= \sum_{x=r}^{n}\left[n\binom{n-1}{x-1}p_0^{x-1}(1-p_0)^{n-x} - n\binom{n-1}{x}p_0^{x}(1-p_0)^{n-x-1}\right] = \\ &= n\binom{n-1}{r-1}p_0^{r-1}(1-p_0)^{n-r}\end{aligned}$$

This term is then integrated with respect to p_0, which means the p-value can be denoted as

$$\begin{aligned}\mathrm{P}[R \geq r|n,p_0] &= n\binom{n-1}{r-1}\int_0^{p_0} p^{r-1}(1-p)^{n-r}dp = \\ &= \frac{1}{B(r,n-r+1)}\int_0^{p_0} p^{r-1}(1-p)^{n-r}dp. \qquad (3.19)\end{aligned}$$

The conversion of the p-value allows to develop a quantified relationship to the posterior probability. The starting point for that is the reformulation of the posterior probability (Equation 3.15) into $\mathrm{P}[p \leq p_0|n,r] = 1 - \mathrm{P}[p > p_0|n,r]$ from which one can conclude that for

the prior choice of $a = 0$ and $b = 1$, the p-value is the same as $\mathrm{P}[p \leq p_0|n, r]$. The latter can be retraced via the formulas

$$\begin{aligned} \mathrm{P}[p \leq p_0|n, r] &= \frac{1}{B(a + r, b + n - r)} \int_0^{p_0} p^{a+r-1}(1 - p)^{b+n-r-1} dp \quad \overset{\substack{a=0 \\ b=1}}{=} \\ &= \frac{1}{B(r, n - r + 1)} \int_0^{p_0} p^{r-1}(1 - p)^{n-r} dp \quad \overset{3.19}{=} \mathrm{P}[R \geq r|n, p_0] \end{aligned} \tag{3.20}$$

It follows from Equation 3.13 that $\mathrm{P}[p \leq p_0|n, r] \leq \alpha$ is the frequentist decision rule which is equivalent to $1 - \mathrm{P}[p > p_0|n, r] \leq \alpha$ and to $1 - \alpha \leq \mathrm{P}[p > p_0|n, r]$. Therefore, choosing $\tau = 1 - \alpha$ as threshold for the posterior probability rule results in the same T1E control for the Bayesian decision as for the binomial test, when the prior $Beta(0, 1)$ is used. This holds for all n, r, and p_0 as can be seen in Equation 3.20.

Next it is of interest whether the connection between the Bayesian decision and the binomial test can be quantified under a different prior, namely the uniform prior defined as $Beta(1, 1)$. For this purpose $\mathcal{P}$ and $\mathcal{B}$ are introduced to simplify the notation of the p-value and the posterior probability using a uniform prior

$$\begin{aligned} \mathcal{P} &:= \frac{1}{B(r, n - r + 1)} \int_0^{p_0} p^{r-1}(1 - p)^{n-r} dp, \\ \mathcal{B} &:= \frac{1}{B(r + 1, n - r + 1)} \int_0^{p_0} p^{r}(1 - p)^{n-r} dp. \end{aligned}$$

From Zaslavsky (2010) it is known that, with a uniform prior, $\mathcal{B} < \mathcal{P}$ holds true. Hence, it follows that τ must at least fulfill $\tau > 1-\alpha$ to enable the same decisions because the binomial test rejects H_0 if $\mathcal{P} \leq \alpha$. Since $\mathcal{B} < \mathcal{P}$, it follows that $1 - \mathcal{B}$ must at least exceed $1 - \alpha$. However, this is the most general condition, as $1 - \alpha$ is the lower boundary for the space of τ. This is thus a necessary but not sufficient condition which can be seen in Figure 3.4 where similar T1E control is only possible for some n, but in the majority, the Bayesian decisions are overoptimistic which results in higher T1E.

To better deal with this issue, a quantification of the difference between $\mathcal{P}$ and $\mathcal{B}$ is elaborated. The difference is considered as a factor which allows the multiplicative conversion of $\mathcal{B}$ to $\mathcal{P}$ and vice versa. The factor has moreover the practical advantage, that the posterior probability can be corrected and then be compared to a constant threshold τ^*. The correction

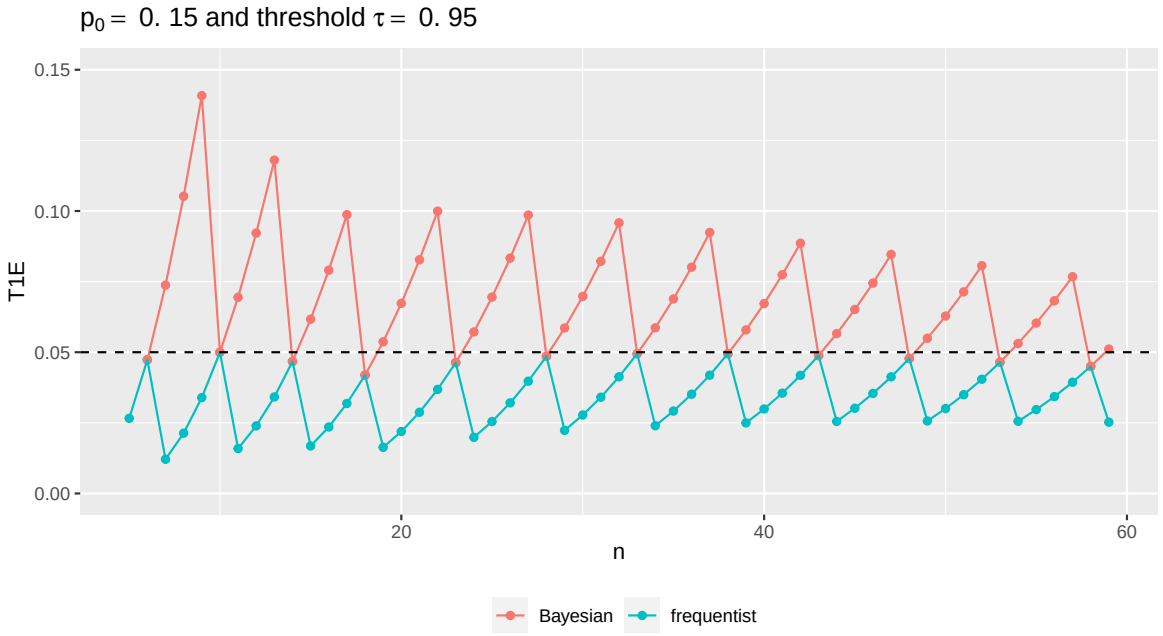


Figure 3.4: *Type 1 error of a one-sided frequentist binomial test with* $\alpha = 0.05$ *and Bayesian decision from a beta-binomial model with uniform prior* $a = b = 1$ *showing that* $\tau = 1 - \alpha$ *is only necessary but not sufficient and reflects the lower boundary for the space of* τ*.*

factor is denoted as κ and without loss of generality

$$\mathcal{P} := \kappa \cdot \mathcal{B} \tag{3.21}$$

which means, in combination with the knowledge of $\mathcal{B} < \mathcal{P}$, that κ will be larger than 1. The correction factor has the following formula

$$\kappa := \frac{\mathcal{P}}{\mathcal{B}} = \frac{\frac{1}{B(r, n-r+1)} \int_0^{p_0} p^{r-1}(1-p)^{n-r} dp}{\frac{1}{B(r+1, n-r+1)} \int_0^{p_0} p^{r}(1-p)^{n-r} dp} =$$

$$= \frac{B(r+1, n-r+1)}{B(r, n-r+1)} \cdot \frac{\int_0^{p_0} p^{r-1}(1-p)^{n-r} dp}{\int_0^{p_0} p^{r}(1-p)^{n-r} dp}.$$

The factor κ consists of two elements. The first is the fraction of two beta functions, and the second is a fraction of two integrals. The first element of κ is investigated and simplified

using the connection of the beta function with the gamma function and results in

$$\begin{aligned}\frac{B(r+1,n-r+1)}{B(r,n-r+1)} &= \frac{\frac{\Gamma(r+1)\cdot\Gamma(n-r+1)}{\Gamma(r+1+n-r+1)}}{\frac{\Gamma(r)\cdot\Gamma(n-r+1)}{\Gamma(r+n-r+1)}} = \\ &= \frac{\Gamma(r+1)\cdot\Gamma(n-r+1)}{\Gamma(r+1+n-r+1)}\cdot\frac{\Gamma(r+n-r+1)}{\Gamma(r)\cdot\Gamma(n-r+1)} = \\ &= \frac{\Gamma(r+1)\cdot\Gamma(n-r+1)}{\Gamma(n+2)}\cdot\frac{\Gamma(n+1)}{\Gamma(r)\cdot\Gamma(n-r+1)} = \\ &= \frac{\Gamma(r+1)}{\Gamma(n+2)}\cdot\frac{\Gamma(n+1)}{\Gamma(r)} = \frac{r\cdot\Gamma(r)}{(n+1)\cdot\Gamma(n+1)}\cdot\frac{\Gamma(n+1)}{\Gamma(r)} = \frac{r}{n+1} \end{aligned} \tag{3.22}$$

meaning that the first fraction is smaller than 1 because of $r \in \{0, ..., n\}$.

For the assessment of the second element the notation $\mathcal{P}'$ and $\mathcal{B}'$ for the elements of the integral quotient is introduced

$$\frac{\mathcal{P}'}{\mathcal{B}'} := \frac{\int_0^{p_0} p^{r-1}(1-p)^{n-r}dp}{\int_0^{p_0} p^{r}(1-p)^{n-r}dp}. \tag{3.23}$$

Both $\mathcal{P}'$ and $\mathcal{B}'$ have a similar structure and the only difference is the power of p in the integrand. Therefore, $\mathcal{B}'$ is investigated applying partial integration ($\int g(x)h'(x)dx = g(x)h(x) - \int g'(x)h(x)dx$) which results in

$$\begin{aligned}\mathcal{B}' = \int_0^{p_0} p^r(1-p)^{n-r}dp &= \left[\frac{p^r(1-p)^{n-r+1}(-1)}{n-r+1}\right]_0^{p_0} - \int_0^{p_0}\frac{rp^{r-1}(1-p)^{n-r+1}(-1)}{n-r+1}dp = \\ &= -\frac{p_0^r(1-p_0)^{n-r+1}}{n-r+1} - \frac{(-1)r}{n-r+1}\int_0^{p_0} p^{r-1}(1-p)^{n-r+1}dp = \\ &= -c_1 + c_2\int_0^{p_0} p^{r-1}(1-p)^{n-r}(1-p)dp = \\ &= -c_1 + c_2\int_0^{p_0} p^{r-1}(1-p)^{n-r} - p^r(1-p)^{n-r}dp = \\ &= -c_1 + c_2\left[\int_0^{p_0} p^{r-1}(1-p)^{n-r}dp - \int_0^{p_0} p^r(1-p)^{n-r}dp\right] = \\ &= -c_1 + c_2\left(\mathcal{P}' - \mathcal{B}'\right)\end{aligned}$$

Now this term is solved with respect to $\mathcal{P}'$

$$\begin{aligned}
\mathcal{B}' &= -c_1 + c_2\left(\mathcal{P}' - \mathcal{B}'\right) &\iff\\
\mathcal{B}' &= -c_1 + c_2\mathcal{P}' - c_2\mathcal{B}' &\iff\\
(1+c_2)\cdot\mathcal{B}' &= -c_1 + c_2\mathcal{P}' &\iff\\
(1+c_2)\cdot\mathcal{B}' + c_1 &= c_2\mathcal{P}' &\iff\\
\mathcal{P}' &= \frac{c_2}{c_2}\mathcal{B}' + \frac{\mathcal{B}'}{c_2} + \frac{c_1}{c_2} = \mathcal{B}' + \frac{\mathcal{B}'}{c_2} + \frac{c_1}{c_2} && (3.24)
\end{aligned}$$

The solved $\mathcal{P}'$ is then incorporated into Equation 3.23 and combined with the simplified fraction of the beta functions in Equation 3.22 to calculate the correction factor κ as

$$\begin{aligned}
\kappa = \frac{\mathcal{P}}{\mathcal{B}} &= \frac{r}{n+1}\cdot\frac{\mathcal{P}'}{\mathcal{B}'}\frac{r}{n+1}\cdot\frac{\left(\mathcal{B}' + \frac{\mathcal{B}'}{c_2} + \frac{c_1}{c_2}\right)}{\mathcal{B}'}\\
&= \frac{r}{n+1}\cdot\left(1 + \frac{1}{c_2} + \frac{c_1}{\mathcal{B}'\cdot c_2}\right)\\
&= \frac{r}{n+1} + \left(\frac{r}{n+1}\frac{1}{\frac{r}{n-r+1}}\right) + \left(\frac{r}{n+1}\cdot\frac{\frac{p_0^r(1-p_0)^{n-r+1}}{n-r+1}}{\mathcal{B}'\cdot\frac{r}{n-r+1}}\right)\\
&= \frac{r}{n+1} + \frac{n-r+1}{n+1} + \frac{r}{n+1}\cdot\frac{p_0^r(1-p_0)^{n-r+1}}{\mathcal{B}'\cdot r}\\
&= \frac{n+1}{n+1} + \frac{p_0^r(1-p_0)^{n-r+1}}{\mathcal{B}'\cdot(n+1)}\\
&= 1 + \frac{p_0^r(1-p_0)^{n-r+1}}{(n+1)\int_0^{p_0} p^r(1-p)^{n-r}dp} =: \kappa(n,r,p_0). && (3.25)
\end{aligned}$$

The correction factor now allows to directly calculate the frequentist p-value ($\mathcal{P}$) from the Bayesian posterior probability and vice versa. Consequently, the p-value and the posterior probability can be considered as functions of each other.

Moreover, a dynamic τ dependent on the known and observed values n, r, p_0 can be determined for the underlying data situation. For the frequentist decision rule it is given that $\mathcal{P} \leq \alpha$ must be fulfilled to decide in favour of the alternative hypothesis. Based on Equation

3.21, the condition is rewritten as $\kappa \cdot \mathcal{B} \leq \alpha$ which is equivalent to $\mathcal{B} \leq \frac{\alpha}{\kappa}$ from which follows

$$1 - \mathcal{B} \geq 1 - \frac{\alpha}{\kappa}$$

and, consequently, the dynamic threshold τ can be seen in dependence of n, r, p_0 in form of

$$\tau_{(n,r,p_0)} = 1 - \frac{\alpha}{\kappa(n, r, p_0)} \tag{3.26}$$

Moreover, the correction factor κ is larger than 1 because the second term in Equation 3.25 is larger than 0, as the reference value p_0 lies in the interval $(0, 1)$. Consequently, Equation 3.25 proves that $\mathcal{B} < \mathcal{P}$ under a uniform prior distribution. This characteristic has been shown before by Zaslavsky (2010) and was also mentioned in this chapter to grasp the lower bound for τ. However, this characteristic was not used to elaborate κ and therefore, in the end, the elaborated κ proves the finding of Zaslavsky (2010) in a different way. The innovation is that κ quantifies the known gap between the frequentist and Bayesian tool in form of a multiplicative factor.

The correction factor is displayed in Figure 3.5 for different total numbers of patients. The figure reveals that κ is monotonically increasing in the number of responses and that the maximum at $r = n$ is independent of the number of patients. This characteristic can be proven by setting $r = n$ into Equation 3.25 which results in

$$1 + \frac{p_0^n(1-p_0)^{n-n+1}}{(n+1)\int_0^{p_0} p^n(1-p)^{n-n}dp} = 1 + \frac{p_0^n(1-p_0)}{(n+1)\left[\frac{p^{n+1}}{n+1}\right]_0^{p_0}} =$$

$$1 + \frac{p_0^n(1-p_0)}{(n+1)\left(\frac{p_0^{n+1}}{n+1} - 0\right)} = 1 + \frac{p_0^n(1-p_0)}{p_0^{n+1}} = 1 + \frac{p_0^n - p_0^{n+1}}{p_0^{n+1}} = 1 + \frac{1}{p_0} - 1 = \frac{1}{p_0} := \kappa_{max},$$

which proves that the correction factor κ increases as the reference value p_0 declines (cf. Appendix A.1 with additional plots under different p_0).

Figure 3.5 also shows that κ remains flat until a certain number of responses has been reached. After this point has been reached, κ starts to rise and converts into a linear increase. The

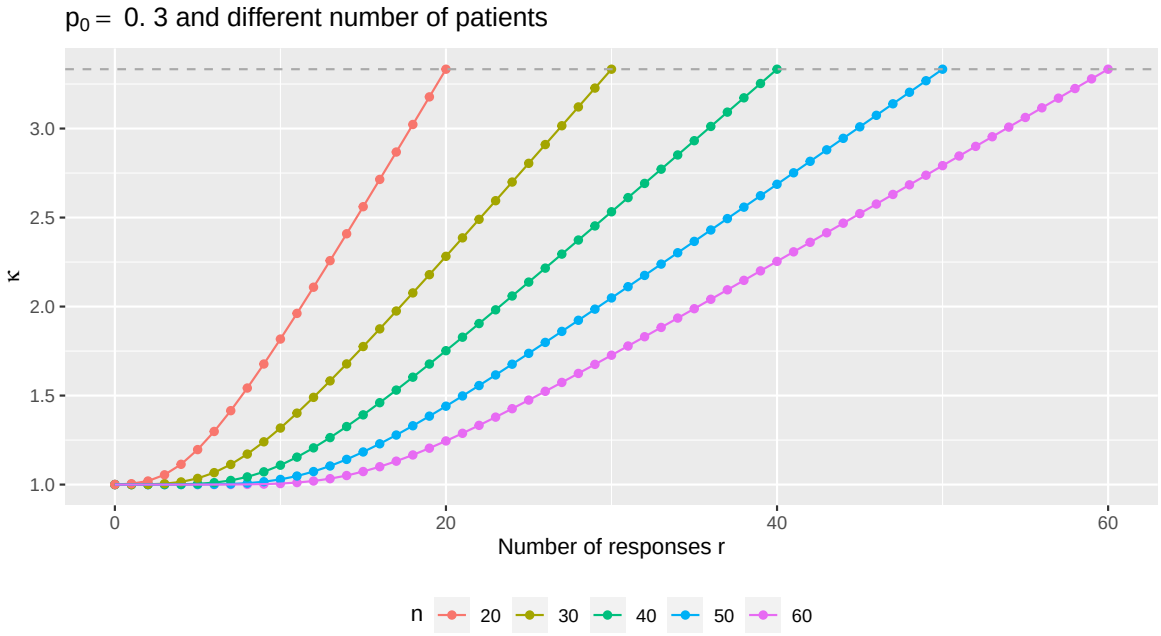


Figure 3.5: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.30$.*

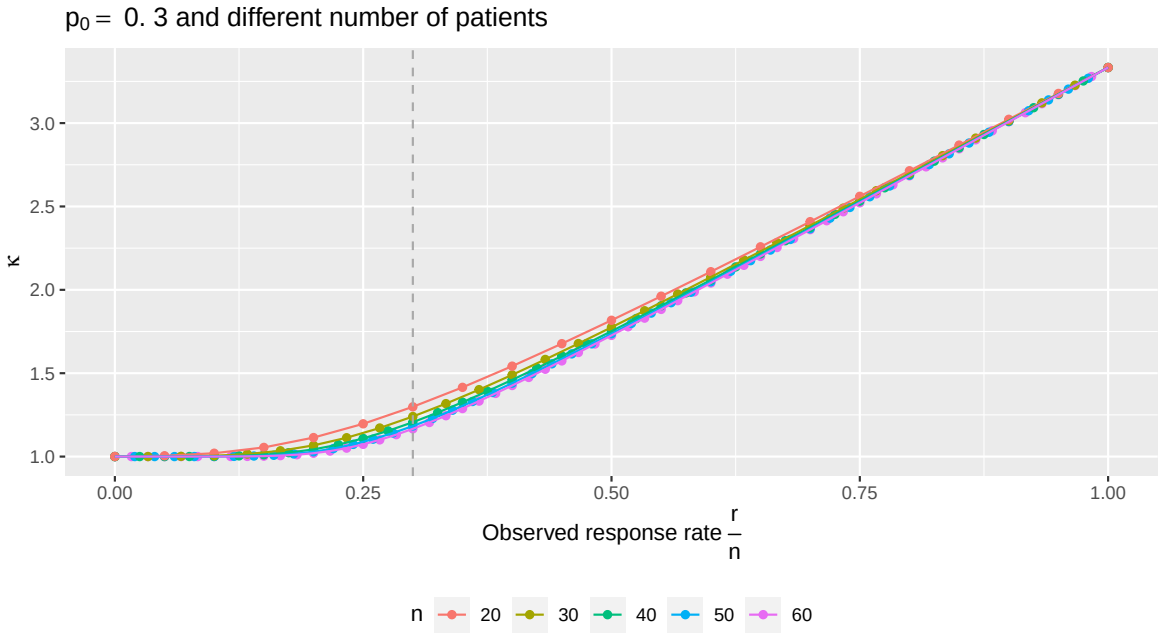


Figure 3.6: *Correction factor κ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value $p_0 = 0.30$.*

next step is to get a closer look at the area of r in which the first rise is observed. At first glance, it seems that κ increases the closer the observed response rate $\frac{r}{n}$ comes to p_0. Hence, κ will be displayed in relation to the observed response rate instead of only r. Figure 3.6 underlines what was hypothesized before. The closer the observed response rate is to p_0, the larger κ becomes. The consequence is a substantially increased κ as regions of r are reached in which the frequentist test starts to reject the null hypothesis. This means that the Bayesian posterior rule is too optimistic and decides too early in favour of the alternative hypothesis if the threshold τ is chosen naively. This overoptimism finally results in a larger T1E. Hence, it is important to know by how much exactly $\mathcal{B}$ is smaller than $\mathcal{P}$ because the factor has its drawbacks when the absolute values are small. In those cases, the absolute difference can facilitate the conversions between $\mathcal{B}$ and $\mathcal{P}$ and might be helpful to grasp the difference in more detail. The absolute difference is worked out with the elaborated knowledge from the ratio. As of Equation 3.24, it is possible to display $\mathcal{P}'$ as a function of $\mathcal{B}'$, and, because of the definition of $\mathcal{P} = \frac{1}{B(r, n-r+1)} \cdot \mathcal{P}'$, it follows that the p-value is given by

$$\mathcal{P} = \frac{1}{B(r, n-r+1)} \left(\mathcal{B}' + \frac{\mathcal{B}'}{c_2} + \frac{c_1}{c_2} \right),$$

and with the property of the beta function $B(r, n-r+1) = \frac{n+1}{r} \cdot B(r+1, n-r+1)$ and with $c_1 = \frac{p_0^r (1-p_0)^{n-r+1}}{n-r+1}$, $c_2 = \frac{r}{n-r+1}$, it further follows that $\mathcal{P}$ is a function of $\mathcal{B}$

$$\begin{aligned}
\mathcal{P} &= \frac{1}{\frac{n+1}{r} \cdot B(r+1, n-r+1)} \left(\mathcal{B}' + \frac{\mathcal{B}'}{c_2} + \frac{c_1}{c_2} \right) = \\
&= \frac{r}{n+1} \cdot \left(\mathcal{B} + \frac{\mathcal{B}}{\frac{r}{n-r+1}} + \frac{1}{B(r+1, n-r+1)} \cdot \frac{p_0^r (1-p_0)^{n-r+1}}{r} \right) \\
&= \frac{r \cdot \mathcal{B}}{n+1} + \frac{(n-r+1) \cdot \mathcal{B}}{n+1} + \frac{p_0^r (1-p_0)^{n-r+1}}{(n+1) \cdot B(r+1, n-r+1)} \\
&= \frac{r \cdot \mathcal{B} + n \cdot \mathcal{B} - r \cdot \mathcal{B} + \mathcal{B}}{n+1} + \frac{p_0^r (1-p_0)^{n-r+1}}{(n+1) \cdot B(r+1, n-r+1)} \\
&= \mathcal{B} + \frac{p_0^r (1-p_0)^{n-r+1}}{(n+1) \cdot B(r+1, n-r+1)}.
\end{aligned}$$

This consequently leads to the analytically quantifiable difference between the Bayesian posterior probability in the beta-binomial model with uniform prior and the p-value of the one-sided binomial test, with the difference being

$$\delta := \mathcal{P} - \mathcal{B} = \frac{p_0^r(1-p_0)^{n-r+1}}{(n+1)\cdot B(r+1, n-r+1)}. \tag{3.27}$$

In Figure 3.7, the absolute difference δ is displayed, showing that it can take values which are of importance when decisions are made based on $\mathcal{B}$. For example for $n = 30$, the difference δ can be more than 10 percentage points. Also the asymptotic characteristic of the difference becomes visible. For an increasing total number of observations n, the absolute difference in general declines and converges to 0 because a constant which is smaller than 1 with the power n declines faster than the multiplication with n increases. It therefore holds for the absolute difference δ that

$$\begin{aligned}\delta &= \frac{p_0^r(1-p_0)^{n-r+1}}{(n+1)} \cdot \frac{\Gamma(n+2)}{\Gamma(r+1)\Gamma(n-r+1)} = \\ &= \frac{p_0^r(1-p_0)^{n-r+1}}{(n+1)\Gamma(r+1)} \cdot \frac{(n+2)!}{(n-r+1)!} \xrightarrow{n\to\infty} 0.\end{aligned}$$

The magnitude of the absolute difference δ in regions where discriminatory decisions with respect to the hypothesis are made is high (see Figure 3.8). This corresponds to the results for the correction factor κ. Additional plots of δ in relation to the observed response rate for different reference values p_0 are given in Appendix A.2 and underline what was observed for the factor; a general decline of the absolute difference with increasing p_0 and also n.

The elaborated analytical relationship allows to interchangeably apply the frequentist binomial test and decisions based on the posterior distribution from a beta-binomial model with uniform prior. The dynamic threshold ensures equal decisions because a static threshold cannot be applied under all scenarios. The correction factor and the absolute difference quantify by how much the decision measures, p-value and posterior probability, differ and they therefore underline how much adaption is required to come to the same decision.

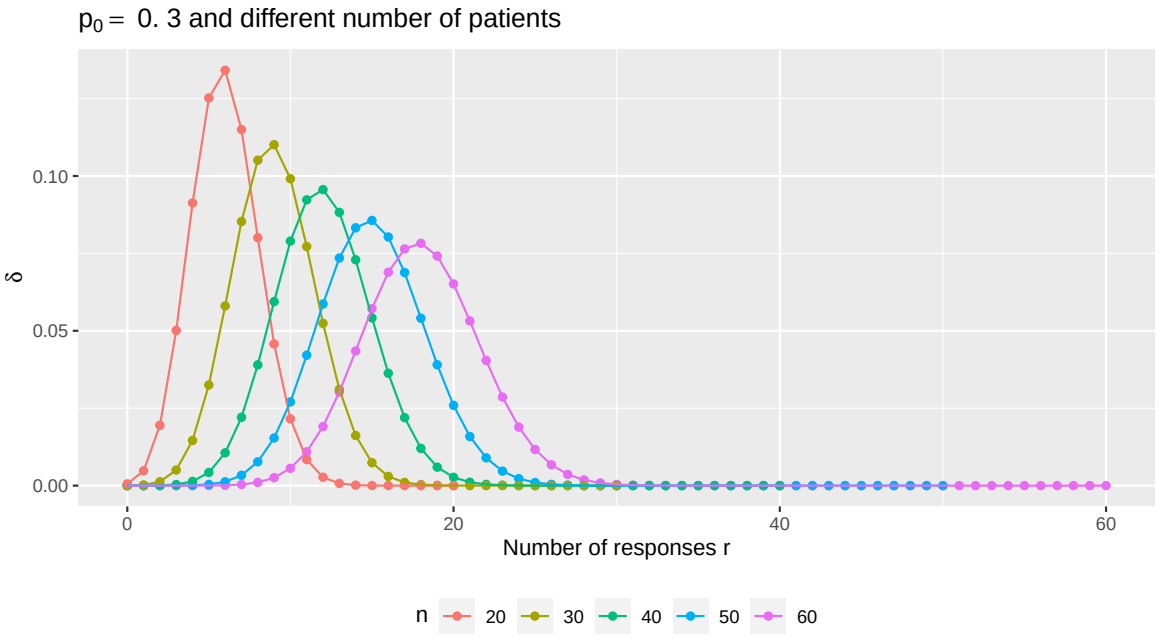


Figure 3.7: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed number of responses r.*

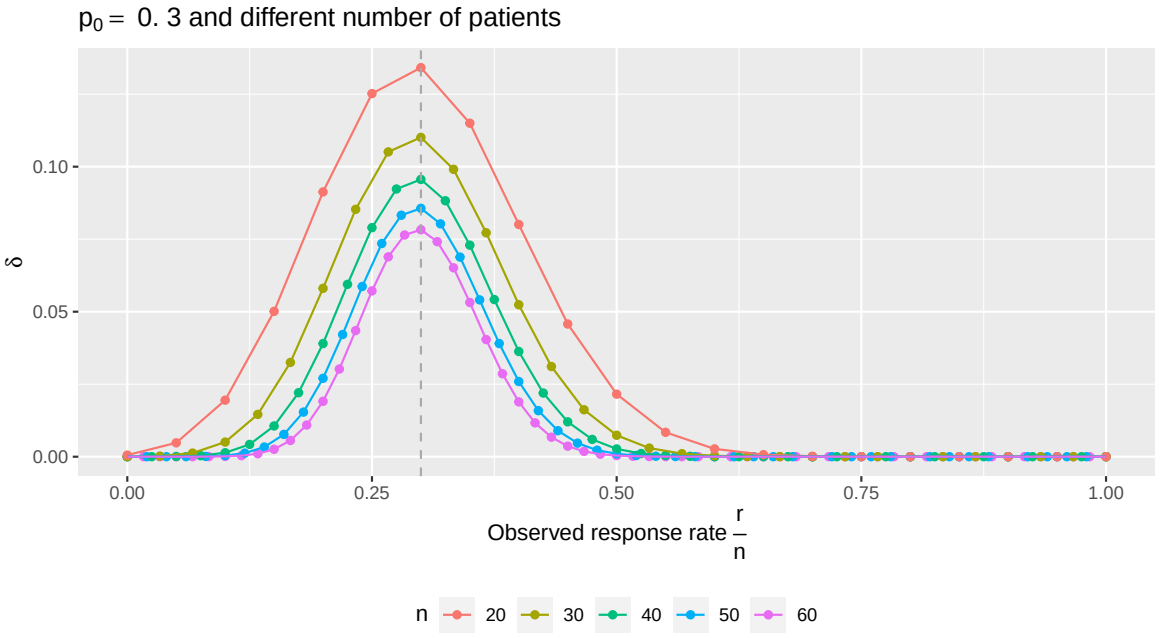


Figure 3.8: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value $p_0 = 0.30$.*

3.2.2 Conversion of a frequentist basket trial into a design with Bayesian tools

In this section, the elaborated connection between the frequentist binomial test and the Bayesian posterior probability is applied. The frequentist design of Cunanan et al. (2017b) will be modified to a basket trial design that applies Bayesian techniques. This will show that tools can be interchanged and, although they appear completely different, the same decisions are made. This also means the same erroneous decisions are made and therefore the T1E remains untouched. The design of Cunanan et al. (2017b) and its tools are described in Section 2.1.2.

The specific setting for a basket trial is taken from the simulation study in Cunanan et al. (2017b), where the authors compare their proposed design with a reference design. For their trial setting, they oriented themselves on talks with experts and on the reported characteristics of the applied basket trial by Hyman et al. (2015). Therefore, their example works with $I = 5$ baskets. For the interim assessment, they require at least $r_s \geq 1$ responses in the heterogeneous and $r_c \geq 5$ responses in the homogeneous path to continue the individual basket and the pooled trial, respectively. In the first stage $n_1 = 35$ patients are recruited. In general, the authors assume an equal accrual of patients, meaning that on average 7 patients per basket are available after stage 1. In the second stage, additional $n_2 = 20$ patients are added in the homogeneous path, while in the heterogeneous path n_{2i} additional patients are recruited to each basket that passed the first stage. The authors simulated n_{2i} together with the significance level for the final analysis in the homogeneous path α_c, the significance level in the heterogeneous path before correction for multiplicity α_s, and the tuning parameter γ in the assessment of heterogeneity. The premise for the simulation was to reach the same family-wise T1E as the reference design under the global null hypothesis (all baskets are futile). The simulation returned $n_{2i} = 15$, $\alpha_c = 0.05$, $\alpha_s = 0.07$, and $\gamma = 0.52$.

With all this information, the frequentist assessments of futility after stage 1 and the final efficacy evaluation after stage 2 are converted to Bayesian decision rules. The null hypothesis assumes a response rate lower than or equal to $p_0 = 0.15$ ($H_0 : p \leq p_0$).

Stage 1: Futility decisions after in total $n_1 = 35$ patients are recruited

- Homogeneous path: The example requires at least $r_c = 5$ responses within the pooled baskets at the interim futility assessment. In a statistical test setting, this corresponds to a significance level of at least $\mathrm{P}[X \geq 5|p = 0.15, n = 35] = 0.6193$ for the one-sided binomial test. The upper boundary is 0.7912, the p-value for $r = 4$ which is the highest number of responses that results in a futility stop. Consequently, the trial continues for any significance level within $[0.6193, 0.7912)$. The values within the interval are rather uncommon choices for the significance level in a statistical test, however, it reflects the liberal layout of the interim futility assessment with the purpose to continue the trial and stop only if barely any support for a potentially efficacious treatment is given. The posterior probability for the Bayesian futility assessment in this setting is derived using the p-value and the elaborated absolute difference δ (Equation 3.27)

$$\mathrm{P}[p \leq p_0|n = 35, r = 5] = 0.6193 - \frac{0.15^5(1 - 0.15)^{35-7+1}}{(35 + 1) \cdot B(5 + 1, 35 - 5 + 1)} = \\ = 0.6193 - 0.1599 = 0.4594.$$

  Consequently the posterior probability to assess interim futility is

$$\mathrm{P}[p > p_0|n = 35, r = 5] = 1 - 0.4594 = 0.5406.$$

  The threshold for the interim futility assessment τ_F can then be derived because it is known that the trial must continue if the posterior probability is 0.5406 or higher (cf. Equation 3.1). Therefore, the trial is stopped for any τ_F in $[0.3550, 0.5406)$ because the lower bound of the interval is the posterior probability for $r = 4$.

- Heterogeneous path: At least $r_c = 1$ responses among the expected $n_{1i} = 7$ patients within each basket are required. Similarly to the homogeneous path, this requirement can be transposed into a statistical test setting for a one-sided binomial test that results in a significance level of at least $\mathrm{P}[X \geq 1|p = 0.15, n = 7] = 0.6794$ and an upper boundary 1, which is the p-value for $r = 0$. Hence, a significance level from the interval $[0.6794, 1)$ ensures that the basket continues only if at least one response is observed.

The corresponding posterior probability to assess interim futility is then derived via

$$\mathrm{P}[p \leq p_0 | n = 7, r = 1] = 0.6794 - \frac{0.15^1(1-0.15)^{7-1+1}}{(7+1) \cdot B(1+1, 7-1+1)} =$$
$$= 0.6794 - 0.3366 = 0.3428$$

and is given by

$$\mathrm{P}[p > p_0 | n = 7, r = 1] = 1 - 0.3428 = 0.6572.$$

The threshold for the interim futility assessment τ_F must be smaller than 0.6572 because $r = 1$ is the smallest response for which the basket moves to the second stage. Hence, any value from the interval $[0.2725, 0.6572)$ may be feasible because the lower boundary is the posterior probability for $r = 0$, which results in a stop.

Stage 2: Efficacy assessment after completed recruitment with all available patients

- Homogeneous path: Additional $n_2 = 20$ patients are recruited, hence, the total number of patients is $n = n_1 + n_2 = 55$. The significance level is defined as $\alpha_c = 0.05$ based on the simulations. Therefore, the smallest number of total responses to reject the null hypothesis with a one-sided binomial test is

$$r_{min} := min\{r : \mathrm{P}[R \geq r | p = 0.15, n = 55] \leq \alpha_c\} = 14$$

The corresponding p-value for $r = 14$ is 0.0297. According to the absolute difference between the p-value and the posterior probability (cf. Equation 3.27), it follows that

$$\mathrm{P}[p \leq p_0 | n = 55, r = 14] = 0.0297 - \frac{0.15^{14}(1-0.15)^{55-14+1}}{(55+1) \cdot B(14+1, 55-14+1)} =$$
$$= 0.0297 - 0.0138 = 0.0159.$$

Consequently, the Bayesian posterior probability to declare the treatment efficacious in the final analysis is

$$\mathrm{P}[p > p_0 | n = 55, r = 14] = 1 - \mathrm{P}[p \leq p_0 | n = 55, r = 14] = 1 - 0.159 = 0.9841$$

and the threshold τ_A could be chosen as 0.9841. However, any value for τ_A within the interval $(0.9657, 0.9841]$ would fulfill the efficacy rule denoted in Equation 3.11 because the lower boundary is the posterior probability for $r = 13$, and with that number of responses the treatment would not be declared efficacious.

- Heterogeneous path: Additional $n_{2i} = 15$ patients are recruited to each basket that passed the first stage, hence the total number of patients is $n = 22$, which includes the expected 7 patients per basket in stage 1. The number of baskets that reach the second stage is denoted by i^* and is used for the Bonferroni correction of the significance level resulting in $\frac{\alpha_s}{i^*}$. The significance level for each basket therefore depends on the number of baskets in the second stage and potential levels for $i^* \in \{1, 2, 3, 4, 5\}$ are

$$\frac{0.07}{i^*} = \quad 0.0700 \quad 0.0350 \quad 0.0233 \quad 0.0175 \quad 0.0140.$$

Consequently, the smallest number of responses to reject the null hypothesis is not unique and depends on i^*

$$r_{min} := min\left\{r : \mathrm{P}[R \geq r | p = 0.15, n = 22] \leq \frac{0.07}{i^*}\right\} = \begin{cases} 7, & \text{if } i^* = 1 \\ 8, & \text{if } i^* \in \{2, 3, 4, 5\} \end{cases}$$

because for $r = 7$ the p-value is 0.0368, while it is 0.0114 for $r = 8$.
It then follows for the posterior probability that with $r = 7$

$$\mathrm{P}[p \leq p_0 | n = 22, r = 7] = 0.0368 - \frac{0.15^7(1 - 0.15)^{22-7+1}}{(22 + 1) \cdot B(7 + 1, 22 - 7 + 1)} = \\ = 0.0368 - 0.0216 = 0.0152$$

and, consequently,

$$\mathrm{P}[p > p_0 | n = 22, r = 7] = 1 - \mathrm{P}[p \leq p_0 | n = 22, r = 7] = 1 - 0.0152 = 0.9848,$$

which means the threshold τ_A for the final analysis with in total $i^* = 1$ baskets can be chosen from $(0.9537, 0.9848]$. The lower boundary is the posterior probability for $r = 6$.

For $r = 8$, the posterior probability is similarly derived from the p-value:

$$\begin{aligned} \mathrm{P}[p \leq p_0 | n = 22, r = 8] &= 0.0114 - \frac{0.15^8 (1 - 0.15)^{22-8+1}}{(22+1) \cdot B(8+1, 22-8+1)} = \\ &= 0.0114 - 0.0072 = 0.0042 \end{aligned}$$

and, consequently,

$$\mathrm{P}[p > p_0 | n = 22, r = 8] = 1 - \mathrm{P}[p \leq p_0 | n = 22, r = 8] = 1 - 0.0042 = 0.9958,$$

which results in the threshold choice $\tau_A \in (0.9848, 0.9958]$ for the individual analysis in each of the $i^* \in \{2, 3, 4, 5\}$ baskets.

The example showed that a transformation of a frequentist design into a design with Bayesian decision rules is feasible. For every frequentist decision, a Bayesian decision rule can be derived with respective threshold values for τ_F and τ_A to ensure the same decisions at the interim futility node and at the final analysis in both the homogeneous and the heterogeneous path of Cunanan et al. (2017b)'s basket trial design.

Equal accrual across the baskets is assumed in the example, hence, on average 7 patients are recruited into each basket during stage 1. In practice however, unequal recruitment is likely even if equal accrual rates are assumed. If the sponsor does not, or cannot, wait until the required 7 patients per basket are recruited, minimum and maximum rules for the number of patients can be set. In their example, Cunanan et al. (2017b) propose at least 3 and a maximum of 10 patients per basket in stage 1, while recruitment continues until in total 35 patients are included. In the homogeneous path, the total number of patients does not change under unequal accrual. Therefore, the Bayesian decision rules do not need adaptions for this scenario. On the other site, in the heterogeneous path, unequal accrual results in baskets with unequal numbers of patients and decisions need to be taken in each basket. More precisely, after stage 1, the number of patients per basket can range from 3 to 10. After the recruitment of additional 15 patients in stage 2, the final basket-wise analysis is then based on 18 to 25 patients. The shifted numbers of patients affect the threshold values for the decisions based on the posterior probability because the transformation of the p-value into a posterior probability depends on the total number of observations n (cf. Equation

3.27). The consequences of the proposed minimum and maximum numbers on the conversion of the frequentist tools into Bayesian decision rules are presented in the following.

In stage 1, the requirement to only continue if at least $r = 1$ response is observed among 3 up to 10 patients results in p-values of the one-sided binomial test that range from 0.3859 for $n = 3$ up to 0.8031 for 10. Correspondingly, the posterior probability $\mathrm{P}[p_i > p_0|n, r = 1]$ decreases from 0.8905 for $n = 3$ to 0.4922 for $n = 10$. To ensure the consistency of decisions, the threshold τ_F must be chosen such that the respective basket is stopped for $r = 0$. The p-value is then 1 for every n, however, the posterior probability $\mathrm{P}[p_i > p_0|n, r = 0]$ ranges from 0.5220 ($n = 3$) to 0.1673 ($n = 10$) which causes an intersection of the posterior probabilities between $r = 1$ and $r = 0$. The posterior probabilities for every $n \in \{3, ..., 10\}$ with $r \in \{0, 1\}$ are given in Table 3.2.

Table 3.2: *Posterior probability $P[p_i > p_0|n, r]$ to assess interim futility under varying number of patients per basket n and with the intention to continue the basket if at least $r = 1$ response was observed.*

	n							
	3	4	5	6	7	8	9	10
$r = 1$	0.8905	0.8352	0.7765	0.7166	0.6572	0.5995	0.5443	0.4922
$r = 0$	0.5220	0.4437	0.3771	0.3206	0.2725	0.2316	0.1969	0.1673

Consequently, a constant τ_F that applies for all scenarios ($n \in \{3, ..., 10\}$) in stage 1 cannot be determined. Locally for either $n \in \{4, ..., 10\}$ or $n \in \{3, ..., 9\}$ solutions are possible, because then the maximum posterior probability for $r = 0$ is smaller than the minimum posterior probability under $r = 1$. The threshold τ_F can then be defined as a value between the maximum under $r = 0$ and the minimum posterior probability under $r = 1$, more specifically $[0.4437, 0.4922)$ for $n \in \{4, ..., 10\}$ and $[0.5220, 0.5443)$ for $n \in \{3, ..., 9\}$.

In stage 2, the unequal accrual in stage 1 results in 18 to 25 patients per basket. The frequentist decisions are made based on p-values that are compared with an adjusted significance level which depends on $i^* = \{1, ..., 5\}$, the number of baskets in stage 2. The transformation under equal accrual already showed that the minimum number of responses r_{min} to declare the basket promising varies under different i^*. Consequently, for unequal accrual the transformation has to account for the two-dimensional set $i^* \times n := \{1, ..., 5\} \times \{18, ...25\}$, which

defines the adjusted significance level $\frac{\alpha_s}{i^*} = (0.0700, 0.0350, 0.0233, 0.0175, 0.0140)$ and the number of patients in the investigated basket. The different minimum number of responses r_{min} for theses scenarios are given in Table 3.3.

Table 3.3: *Minimum number of observed responses r_{min} to declare a basket promising in the final analysis under varying number of patients per basket n and varying adjusted significance level $\frac{\alpha_s}{i^*}$.*

i^*	$\frac{\alpha_s}{i^*}$	n							
		18	19	20	21	22	23	24	25
5	0.0140	7	8	8	8	8	9	9	9
4	0.0175	7	7	8	8	8	8	9	9
3	0.0233	7	7	7	8	8	8	8	9
2	0.0350	7	7	7	7	8	8	8	8
1	0.0700	6	6	6	7	7	7	7	7

This results in an increased complexity in the choices for the threshold τ_A of the posterior probability. A consistent solution among both dimensions is not possible as already shown when only i^* varied, however, local solutions over i^* and n are available. The results for the transposed posterior probability under $i^* = 3$ and adjusted significance level $\frac{0.07}{3} = 0.0233$ are presented in the following. According to Equation 3.27, the p-value of the minimum number of responses is converted into the posterior probability. Apart from that, the posterior probability with one response less than the minimum number or responses is given because in that case, the basket must not be declared promising. These posterior probabilities for $n \in \{18, ..., 25\}$ are shown in Table 3.4. The posterior probabilities for $r_{min} - 1$ and r_{min} do not intersect because the maximum posterior probability for $r_{min} - 1$ is smaller than the minimum under r_{min} ($0.9894 < 0.9917$). Therefore, a local solution for τ_A under $i^* = 3$ over all n is feasible and a specific choice from $(0.9894, 0.9917]$ ensures equal decisions in the final analysis for the posterior probability and the one-sided binomial test under observed unequal accrual.

The respective tables with posterior probabilities $\mathrm{P}[p_i > p_0|n, r]$ for the other i^* are given in Appendix A.3 because the procedure is equivalent and the results are similar. For every i^*, there exists a local solution over the potential number of observations n. However, a solution for τ_A over all i^* is not possible because there is no value that is contained in every interval

Table 3.4: *Posterior probability $P[p_i > p_0|n, r]$ for the final analysis with the intention to declare a basket promising if at least r_{min} responses were observed. The number of patients per basket n among the $i^* = 3$ different baskets in stage 2 is varying.*

	n							
	18	19	20	21	22	23	24	25
r_{min}	0.9959	0.9941	0.9917	0.9970	0.9958	0.9941	0.9920	0.9970
$r_{min} - 1$	0.9837	0.9781	0.9713	0.9886	0.9848	0.9801	0.9745	0.9894

generated by the maximum of the posterior probability under $r_{min} - 1$ and the minimum of the posterior probability under r_{min}. The intervals for all i^* are shown in Table 3.5.

Table 3.5: *Upper and lower boundary of the interval that contains the local solution for τ_A over all n given for every i^*.*

	i^*				
	1	2	3	4	5
upper	0.9679	0.9886	0.9917	0.9941	0.9958
lower	0.9632	0.9848	0.9894	0.9920	0.9941

The presented results show that the complexity of the conversion increases in the case of unequal accrual in the heterogeneous path. The interim futility assessment is more challenging because no unique solution for the Bayesian threshold τ_F can be given for all $n_{1i} \in \{3, ..., 10\}$. However, if 3 or 10 is removed, a solution τ_F can be determined to ensure equal decisions in the frequentist and the Bayesian setting. In the final analysis, a local solution for τ_F can be determined. The solution holds true over all $n \in \{18, ..., 25\}$. The only restriction lies in i^*, the number of baskets in stage 2, which is equivalent to the frequentist trial conduct when the significance level is adapted according to i^*.

This subsection showed that the relationship between the frequentist and Bayesian decision rules can be practically applied and leads to the same decisions at the respective nodes of the basket trial design which was proposed by Cunanan et al. (2017b).

3.2.3 Further connections between interim futility tools

Parts of this Subsection 3.2.3 are already published in the article *Categories, components, and techniques in a modular construction of basket trials for application and further research*

by Pohl et al. (2021). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

In this subsection, further connections among interim futility tools for basket trials are elaborated. The futility tools all have a different way to process the available information and to make a decision. The connections allow to tune the tools such that the same decisions can be achieved, as already shown for the binomial test and the posterior probability. The so far observed information from basket i at the interim assessment is stored in $\mathcal{D} = \{n, r\}$, where the index i is dropped for a more clear presentation.

The starting point for the connection map (Figure 3.9) is the minimum number of responses criterion. This rule requires a minimum number of responses $r \geq r_{min}$ to continue with the basket. The connection to the statistical test ($H_0 : p \leq p_0$, $H_1 : p > p_0$) was already used in the previous subsections and is based on the fixed p-value that is returned for the same data. The p-values which are smaller than the significance level α_F reject the null hypothesis and indicate that there is currently enough support for a non-futile treatment effect. Consequently, for a known number of observations n, the significance level α_F can be chosen such that for at least r_{min} responses the statistical test does not prune the basket. Hence, a minimum number of responses rule can be considered as a primitive one-sided binomial test. The connection goes in both directions because a minimum number of response also reflects a certain significance level and a given significance level represents a certain minimum number of responses. The knowledge about the a smallest number of responses r_{min} can be used in the Bayesian setting. The threshold τ_F for the posterior distribution is then chosen such that it fulfills $\mathrm{P}[p_i > p_0 | n, r_{min}] > \tau_F$ and that at the same time for $r < r_{min}$ the posterior probability is smaller than or equal to τ_F (cf. Equation 3.1). This connection between the frequentist statistical test and the Bayesian posterior probability holds when no information was shared before because with shared information, the threshold τ_F also depends on the observations in the other baskets and on the applied sharing tool.

The Bayesian interim futility tool uses the posterior probability with two different reference values, p_0 (Equation 3.1) and $\frac{p_0+p_1}{2}$ (Equation 3.2). In both tools, the same posterior distribution of response rate p is underlying. Hence, the two tools differ in the investigated intervals on the space of p because of the different reference values. Without loss of general-

ity, it is assumed that $p_0 < p_1$ and that for r_{PP}, the posterior probability does not lead to a futility stop because of $\mathrm{P}[p > \frac{p_0+p_1}{2}|n, r_{PP}] > \tau_F^m$, but for $r < r_{PP}$ the basket is pruned. The resulting question is how to choose τ_F such that the same decision is taken while using p_0 as the reference value. It is known that

$$P\left[p > \frac{p_0 + p_1}{2}\middle| n, r\right] = P\left[p > p_0\middle| n, r\right] - P\left[p \in \left(p_0, \frac{p_0 + p_1}{2}\right]\middle| n, r\right] \leq \tau_F^m$$

and that the same decision is taken if $\mathrm{P}[p > p_0|n, r] \leq \tau_F$ holds at the same time. Hence, from

$$P\left[p > p_0\middle| n, r\right] = P\left[p > \frac{p_0 + p_1}{2}\middle| n, r\right] + P\left[p \in \left(p_0, \frac{p_0 + p_1}{2}\right]\middle| n, r\right] \leq \tau_F$$

it follows that $\tau_F = \tau_F^m + \mathrm{P}[p \in (p_0, \frac{p_0+p_1}{2}]|n, r]$, where the last summand is a constant value. The constant increment to τ_F^m is the probability mass in $(p_0, \frac{p_0+p_1}{2}]$ of the investigated posterior distribution. Therefore, both tools take the same decisions when the reference values and the thresholds are adapted. The adaptions in the reference value and the thresholds are interchangeable because both refer to the same posterior distribution. The posterior distribution can also contain shared information from other baskets.

The MUCE design by Lyu et al. (2020) applies a hypothesis driven approach, where the random variable Z_i indicates at which side of the null reference θ_0 the posterior distribution is mainly located. The futility tool is defined as $\mathrm{P}[Z \geq 0|D] < \tau_F$. It follows from the defined characteristic of Z, that for $Z \geq 0$ the transformed response rate θ is located below θ_0, which means that the futility tool can be rewritten as $\mathrm{P}[p > p_0|D] < \tau_F$ and then conforms with the other posterior probability tools.

The posterior predictive probability (PPP) is a function of the current posterior distribution and the number of remaining observations n_2 (cf. Equation 3.4). It calculates the total probability for a successful final analysis. The PPP futility tool can be applied for every reference value and the threshold value τ_F^P is independent of the current number of observations. This is the major difference to the posterior probability tool where smaller n_1 result in higher variances and therefore more uncertainty, which requires regular adaptions of the threshold τ_F as n_1 changes. Still, the two tools go hand in hand because, under fixed n_1 observations, the threshold τ_F can be adapted such that the decisions with both tools are the same. This

is achieved for $\tau_F := \mathrm{P}[p > p_0 | n_1, r_{PPP} - 1]$, where r_{PPP} is the smallest number of responses for which the PPP tool continues. The PPP can be simulated or analytically calculated. The analytic solution works in Fujikawa et al. (2020) because a beta-binomial model is applied, which means that the posterior distribution is beta-distributed and the number of future responses follows a binomial distribution. The following calculation shows that the posterior predictive probability is then a sum over beta functions. The so far used notation rules are expanded by r^*, the number of responses including the shared information, and c_i, $\tilde{c_i}$ as unspecified constant values. The probability of future responses is

$$\begin{aligned}
&\int_0^1 f(r_i|p) f(p|\mathcal{D}) dp = \\
&= \int_0^1 c_i p^{r_i} (1-p)^{n_2 - r_i} \cdot \frac{1}{B(a + r^*, n_1 - r^* + b)} p^{a + r^* - 1} (1-p)^{n_1 - r^* + b - 1} dp \\
&= \frac{c_i}{B(a + r^*, n_1 - r^* + b)} \cdot \int_0^1 p^{a + r_i + r^* - 1} (1-p)^{b + n_1 + n_2 - r^* - r_i - 1} dp \\
&= \tilde{c_i} \cdot B(a + r_i + r^*, b + n_1 + n_2 - r^* - r_i),
\end{aligned}$$

hence, the posterior predictive probability (PPP) is a sum over beta functions

$$\sum_{\substack{\{r_i\text{: final} \\ \text{analysis successful}\}}} \int_0^1 f(r_i|p_i) f(p_i|\mathcal{D}) dp_i = \sum_{\substack{\{r_i\text{: final} \\ \text{analysis successful}\}}} \tilde{c_i} \cdot B(a + r_i + r^*, b + n_1 + n_2 - r^* - r_i).$$

If the posterior distribution at interim does not have a beta distribution, simulations can be used. The posterior distribution at interim might already contain shared information from other baskets.

The frequentist conditional power is a function of the assumed response rate, the number of remaining observations n_2, and the significance level α in the final analysis. The threshold τ_F can be tuned such that the decisions are congruent to the minimum number of responses r_{min} which are needed to not prune that basket. The conditional power tool is then in line with the statistical test and the minimum number of responses. The conditional power and the posterior predictive probability have the same goal, they both skip ahead to the final analysis and evaluate if it is worth to continue up until there. Apart from that, they highlight the difference of a frequentist (point estimate) and a Bayesian (random variable with distribution) approach on how the response rate is evaluated.

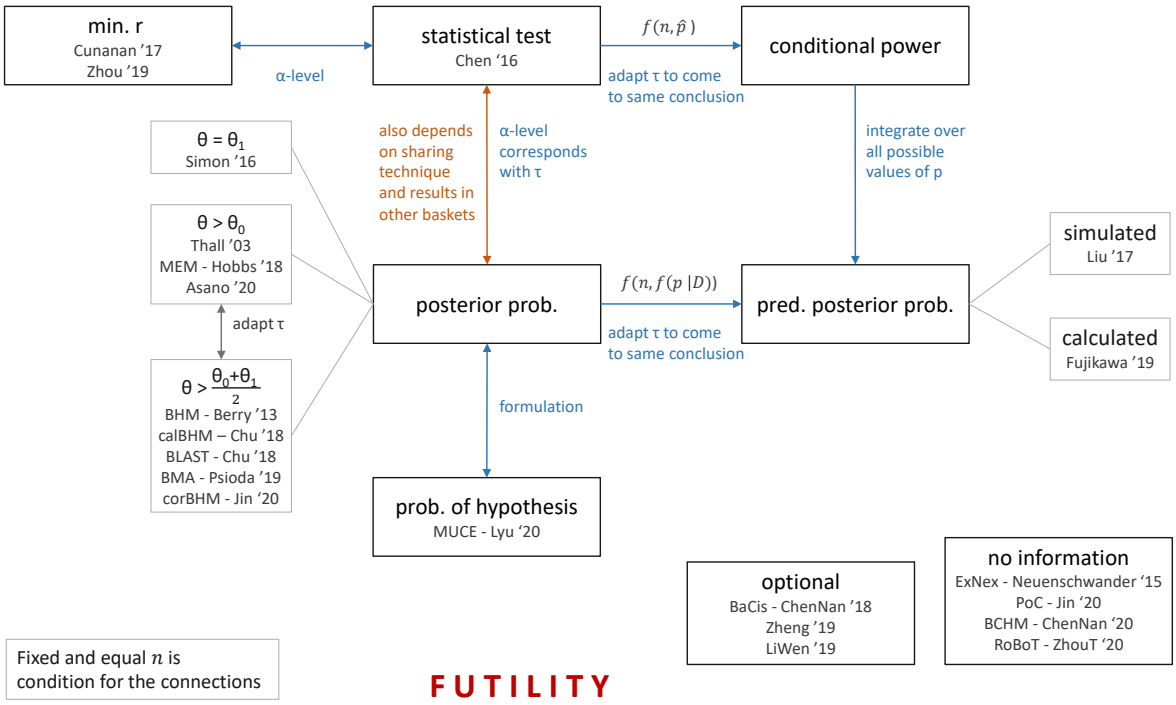


Figure 3.9: *Each box stands for an interim futility tool. The arrows indicate the connections between the tools and the text beside gives further information. The names within the boxes present in which design a tool is applied. Adapted from Pohl et al. (2021).*

The elaborated connections among the interim futility tools are graphically summarized in Figure 3.2. The knowledge about the connection can be helpful when designing a basket trial with the modular framework (cf. Section 3.1). The connections between the futility tools also apply for the interim efficacy tools because of the equivalent structure of the tools.

3.3 Hierarchical beta-binomial model in comparison to BHM for information sharing in basket trials

This section investigates whether a hierarchical beta-binomial model (HBB) is a feasible tool to share information among baskets. Section 3.1.2 shows that the BHM by Berry et al. (2013) is the basic sharing tool which was further developed into many other sharing tools. Also, it shows that non-transformed tools are available. Consequently, the next step is to think of a non-transformed Bayesian hierarchical model as an additional basic tool to share information. In the following, the HBB and its properties are systematically investigated and compared to the BHM.

3.3.1 Comparison of assumed distributions for p

The first aspect of the comparison is to investigate whether and if yes, how the assumed distributions for the response rate p differ. In the BHM, a normal distribution for the logit-transformed response rate, also denoted as a logit-normal distribution ($p \sim logitN(\mu, \sigma^2)$), is assumed, whereas for the HBB a beta distribution ($p \sim beta(a, b)$) is assumed. The difference of a logit-normal distribution and a beta distribution is investigated via the densities. If the distributions were equal, a logit-transformed beta distribution would result in a normal distribution. That assumption holds because a logit-transformed logit-normal distribution has a normal distribution. The logit function is differentiable and non-zero on $(0, 1)$ (see Appendix A.4) and therefore fulfils the requirements for the transformation of a density function. Hence, a logit-transformed beta distribution has the following density function

$$
\begin{aligned}
f_{logitBeta}(x) &= f_{Beta}(logit^{-1}(x)) \cdot \left| \frac{d\ logit^{-1}(x)}{dx} \right| \\
&= \frac{1}{B(a,b)} \cdot (expit(x))^{a-1}(1 - expit(x))^{b-1} \cdot \left| \frac{d\ logit^{-1}(x)}{dx} \right| \\
&= \frac{1}{B(a,b)} \cdot \left(\frac{exp(x)}{1+exp(x)} \right)^{a-1} \left(\frac{1}{1+exp(x)} \right)^{b-1} \cdot \frac{exp(x)(1+exp(x)) - exp(x)^2}{(1+exp(x))^2} \\
&= \frac{1}{B(a,b)} \cdot \left(\frac{exp(x)}{1+exp(x)} \right)^{a-1} \left(\frac{1}{1+exp(x)} \right)^{b-1} \cdot \left(\frac{exp(x)}{1+exp(x)} \right) \left(\frac{1}{1+exp(x)} \right) \\
&= \frac{1}{B(a,b)} \cdot \left(\frac{exp(x)}{1+exp(x)} \right)^{a} \left(\frac{1}{1+exp(x)} \right)^{b}.
\end{aligned}
$$

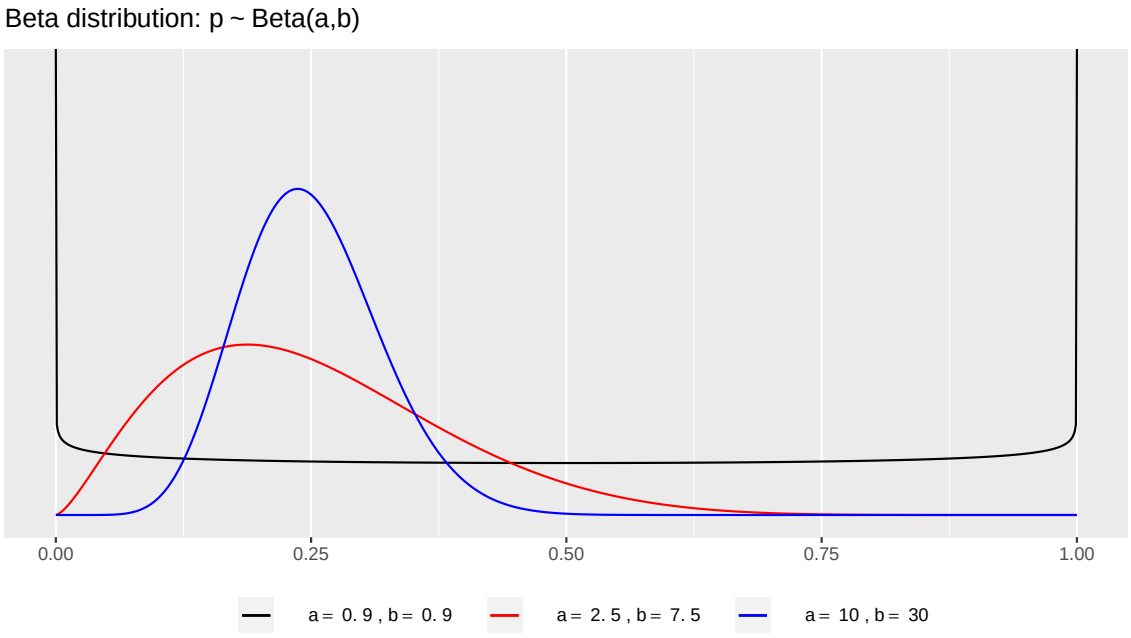


Figure 3.10: *Beta distribution of response rate p with different parameter choices.*

The density does not have the form of a normal distribution. Also it is not symmetric because the mean and the median are different, e.g. for $a = 2, b = 6$ they are -1.2836 and -1.2174. Mean and median were calculated numerically using 10^7 samples. Hence, the logit-transformed beta distribution is not a normal distribution. It then directly follows that an expit-transformed normal distribution is not a beta distribution, because otherwise the logit-transformed beta distribution must have been a normal distribution. This holds because the $expit(x)$ function is the inverse function of $logit(x)$. Finally, this means that the assumed distributions for the response rate p are different in BHM and HBB.

Nevertheless, the logit-normal distribution and the beta distribution have similar characteristics. Both are defined on the interval $(0, 1)$ and are either unimodal or bimodal. Both are not symmetric, except for the parameter choices $\mu = 0$ and $a = b$. Different behaviour is observed at the interval borders. The logit-normal distribution converges to 0 whereas the beta distributions converges to ∞. Figures 3.10 and 3.11 display these characteristics.

In the next step, the aim is to approach one given distribution as close as possible with the other distribution. The procedure is as follows. First, the given distribution is defined. Without loss of generality, the procedure is described for a given beta distribution with

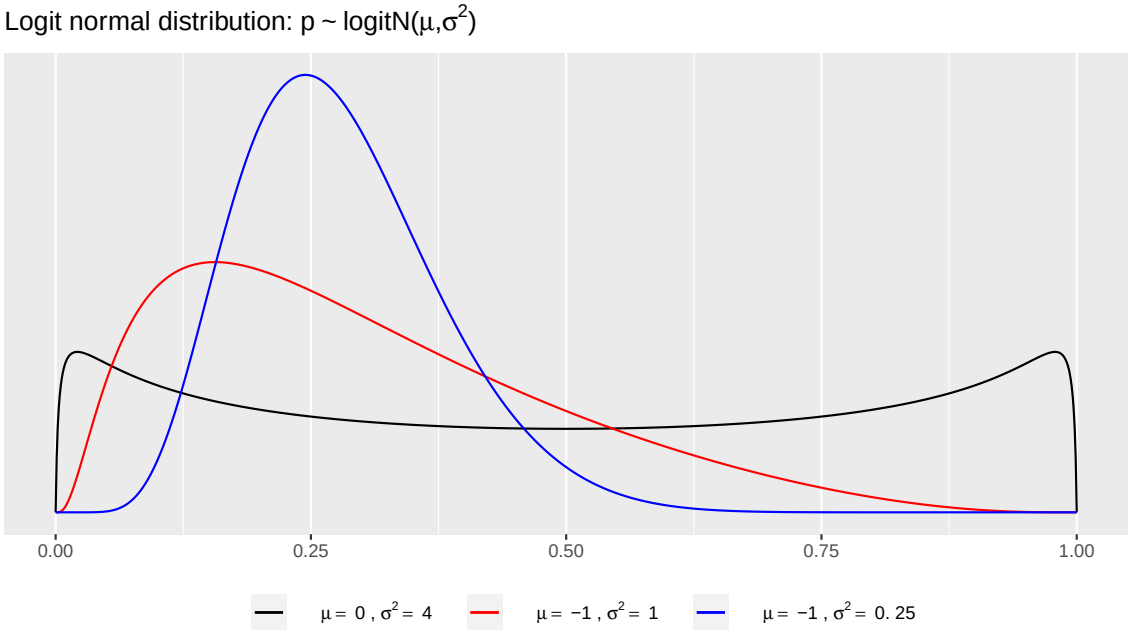


Figure 3.11: *Logit-normal distribution of response rate p with different parameter choices.*

$a = 2$ and $b = 5$. Based on this distribution, a certain number of samples (e.g. 10^5) are drawn and then logit-transformed. These transformed samples are then used to fit a normal distribution. The expected value and the standard deviation are estimated with maximum likelihood estimates (MLE). The MLEs then define the logit-normal distribution that most likely approaches the given beta distribution. For $Beta(2, 5)$ the MLEs are $\mu_{MLE} = -1.08$ and $\sigma^2_{MLE} = 0.93^2$. In case the given distribution is logit-normal, the sampled data from the normal distribution is transformed with the $expit(x)$ function. Plots with the given distribution, the distribution based on MLE parameters, and the empirical distribution of the samples are used to graphically evaluate the differences and the behaviour under different parameter scenarios. In Figure 3.12, the beta distribution is the given distribution, while in Figure 3.13 the logit-normal distribution is displayed. The figures underline the different forms of the distributions and show that adaptions of the parameters do not overcome this structural difference, but allow at least an approximation of each other.

Since the beta distribution converges for increasing n, namely $a, b \to \infty$, towards a normal distribution (Moscovich et al., 2016), it follows with the continuous mapping theorem that the logit-transformed beta distribution converges towards a normal distribution and, therefore, the beta distribution converges towards a logit-normal distribution. Consequently,

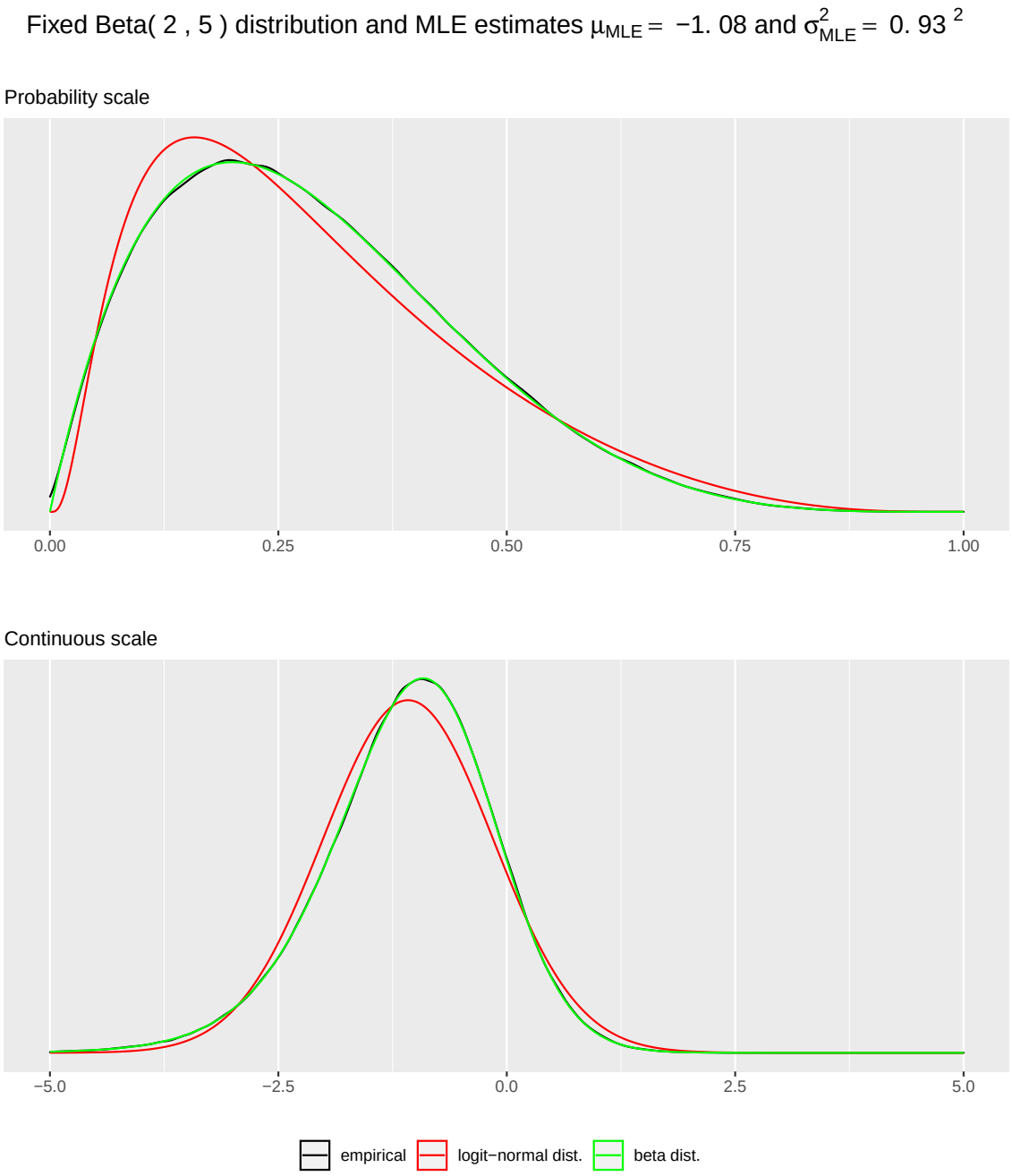


Figure 3.12: *Fixed beta distribution and corresponding maximum likelihood estimated logit-normal distribution.*

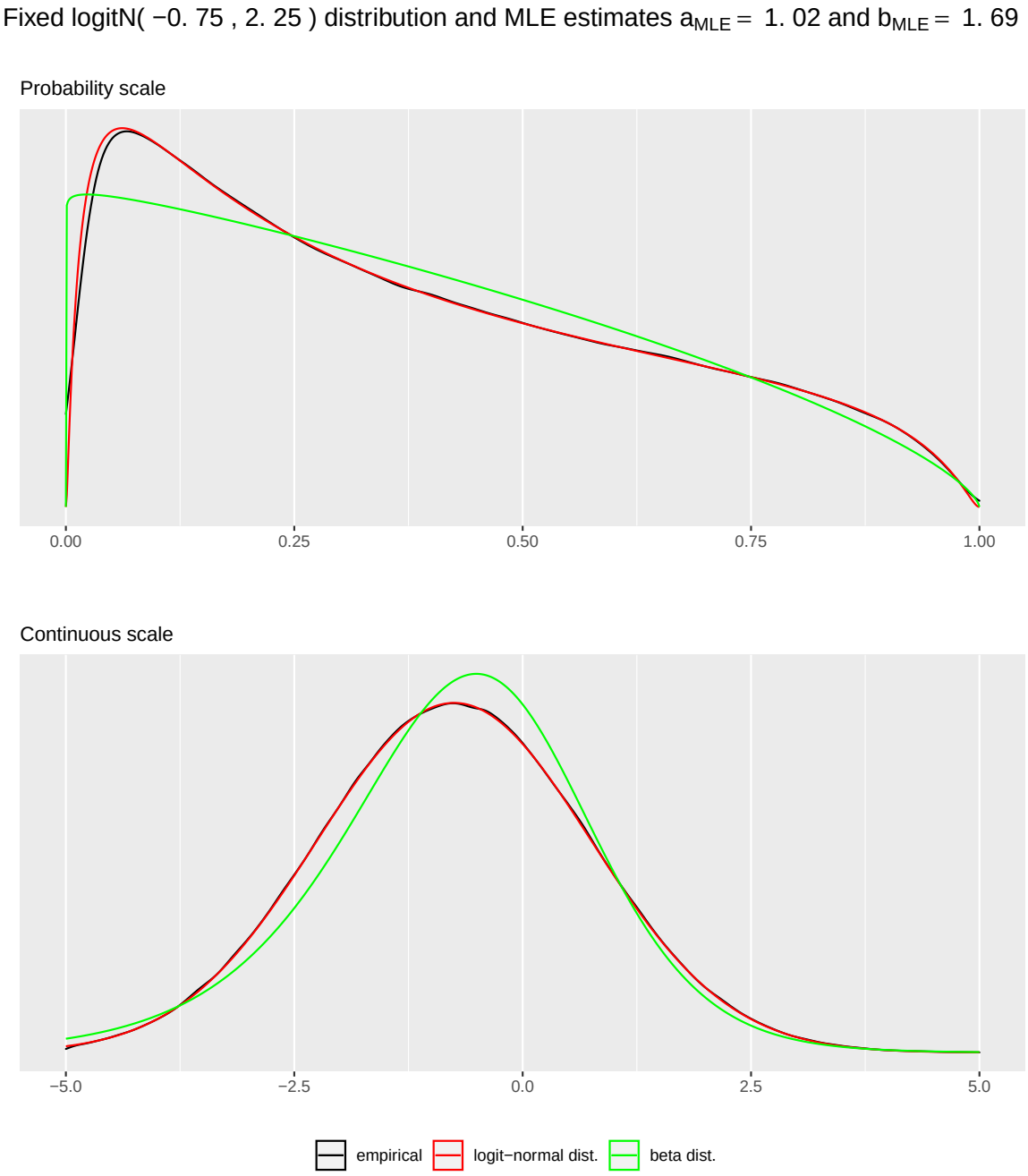


Figure 3.13: *Fixed logit-normal distribution and corresponding maximum likelihood estimated beta distribution.*

asymptotic equivalence can be achieved. However, in applied basket trials the number of observations is often low and the asymptotic characteristic is not reached.

3.3.2 Bayesian behaviour with binomial data

The previous subsection showed that the assumed distributions for the response rate p in BHM and HBB are different. In this subsection the Bayesian behaviour of the two distributions together with binomial data is investigated. Hence, either a logit-normal or a beta prior are assumed, and, together with the binomial likelihood, the respective posterior distributions are calculated. The binomial data is chosen such that it reflects a potential data scenario in one basket in an applied basket trial. In Hyman et al. (2015) for example, half of the baskets contained 10 or less patients at the final analysis. The total number of observations is set to $n = 10$ and the posterior distributions are calculated for all responses scenarios $r \in \{0, 1, 2, ..., 10\}$. The posterior distributions are calculated numerically using the JAGS MCMC sampler via R. The posterior distributions are then compared using their 2.5%, 25%, 50%, 75%, and 97.5% quantiles.

The comparison of the posterior distributions requires that the underlying prior distributions are similar with respect to both their position and to the amount of contained information. The main purpose of the comparison is to investigate how the prior distributions evolve in combination with observed data. Hence, for the prior distribution a prior is used which is as vague as possible. For the logit-normal distribution, this means that the variance σ^2 is large. In orientation to Berry et al. (2013), the prior distribution is then set to $logitN(logit(0.25), 7.8^2)$. This distribution is approximated by $Beta(0.153, 0.183)$, where the parameters are MLEs based on the logit-normal distribution. Both distributions are shown in Figure 3.14.

Tables 3.6 and 3.7 present the quantiles of the posterior distributions. Three posterior distributions are sampled for each response outcome $r \in \{0, 1, 2, ..., 10\}$. The intention behind this is to check the robustness of the numerical calculation. The logit-normal prior distribution results in numerically robust posterior distributions. However, the beta distribution has numerical problems. First, the MCMC sampling in JAGS stops when the chain reaches values for which no density exists, that is for 0 and 1. A pragmatic solution for this problem

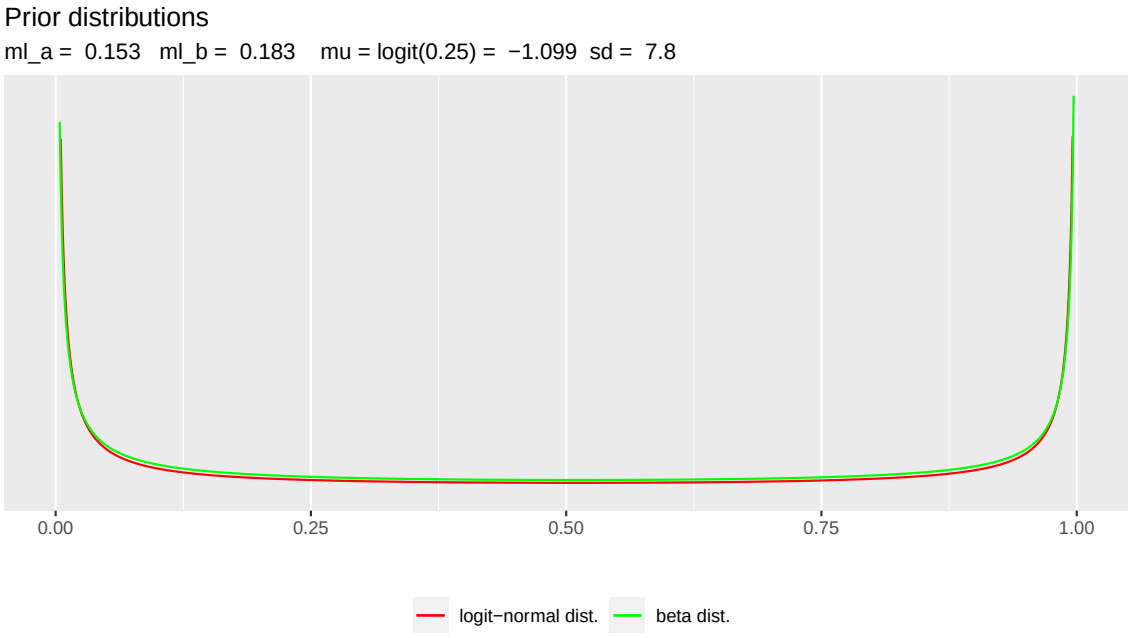


Figure 3.14: *Logit-normal and beta prior distributions to compare their posterior behaviour together with binomial data.*

is to truncate the beta distribution in the JAGS code very close to 0 and 1, e.g. to the interval $(0.000000000001, 0.999999999999999)$ (cf. Appendix B.1). Moreover, the quantiles for $r = 0$ and $r = 10$ are varying, which means that the numerical estimation of the posterior distribution in these scenarios are unstable. The variation can be reduced if the number of MCMC samples is increased. The tables presented here are based on 10^5 MCMC samples. An increase in the number of samples directly corresponds to an increase in the computation time. Despite these drawbacks in the numerical calculation of the posterior distribution for the beta-binomial model, the conjugacy of this prior-likelihood combination allows to calculate the posterior distribution analytically. Hence, the numerical problems can be avoided in practice. Moreover, the numerical instabilities for $r \in \{0, n\}$ reduce for an increasing number of observations n.

The quantiles show that the posterior distributions based on a logit-normal prior (Table 3.6) and based on a the beta prior (Table 3.7) are different and also evolve differently with increasing r. For $r \in \{0, 1, 2, 3\}$, the quantiles of the posterior distribution with logit-normal prior are smaller than those from the beta prior. This means that the logit-normal posterior distribution is rather located towards the left boundary 0 as compared to a beta prior.

Table 3.6: *Quantiles of the posterior distribution and their numerical robustness for every possible number of responses r given the logit-normal prior.*

r	2.5%	25%	50%	75%	97.5%
	0.000000	0.000014	0.000497	0.006977	0.102401
0	0.000000	0.000014	0.000495	0.006992	0.103469
	0.000000	0.000014	0.000489	0.006898	0.102077
	0.003489	0.033816	0.076937	0.145540	0.339187
1	0.003510	0.033487	0.076775	0.145815	0.337552
	0.003503	0.033665	0.076994	0.146281	0.340119
	0.029360	0.108550	0.180779	0.272689	0.481105
2	0.029190	0.108239	0.180462	0.273022	0.481046
	0.029140	0.108144	0.180350	0.272474	0.480833
	0.075262	0.195604	0.286280	0.389760	0.600403
3	0.075390	0.195621	0.286197	0.390404	0.598562
	0.075373	0.195775	0.285874	0.390006	0.598951
	0.136111	0.290017	0.391580	0.500341	0.698575
4	0.137455	0.290409	0.392217	0.500386	0.698509
	0.136319	0.289813	0.391755	0.500247	0.697705
	0.211810	0.391865	0.498313	0.605261	0.785368
5	0.212255	0.390891	0.497951	0.605471	0.785574
	0.212470	0.389919	0.497698	0.605750	0.783773
	0.299233	0.495780	0.603939	0.706214	0.859965
6	0.298270	0.495423	0.604615	0.706633	0.861377
	0.298531	0.495769	0.604706	0.707277	0.861608
	0.395831	0.606021	0.709868	0.801024	0.922537
7	0.395596	0.606409	0.710393	0.801546	0.923380
	0.397012	0.606702	0.710586	0.801170	0.922933
	0.513437	0.722986	0.815698	0.889018	0.970068
8	0.514640	0.722603	0.815411	0.888499	0.969263
	0.513855	0.723040	0.815996	0.888816	0.969640
	0.655639	0.849879	0.919573	0.964039	0.995998
9	0.656514	0.849868	0.919702	0.964270	0.996093
	0.655725	0.849574	0.919406	0.963954	0.996045
	0.884603	0.989801	0.999046	0.999964	1.000000
10	0.884482	0.989777	0.999037	0.999964	1.000000
	0.884056	0.989691	0.999045	0.999963	1.000000

This behaviour starts to change for $r = 4$, where the 75% and 97.5% quantiles for the logit-normal scenario are larger than the respective quantiles from the beta scenario. The same holds for $r = 5$ and for $r = 6$, where all logit-normal quantiles except for the 2.5% quantile are larger. For $r \in \{7, 8, 9, 10\}$ all logit-normal quantiles are larger than the beta quantiles. Consequently, the posterior distribution with logit-normal prior tends towards the boundaries 0 and 1, whereas the posterior distribution with beta prior orientates itself more towards the center. The differences between the quantiles are in particular cases larger than one percentage point, e.g. for $r = 1$, 75% and 97.5%.

Table 3.7: *Quantiles of the posterior distribution and their numerical robustness for every possible number of responses r given the beta prior.*

r	2.5%	25%	50%	75%	97.5%
	0.00000	0.00001	0.00084	0.01173	0.11295
0	0.00000	0.00001	0.00100	0.01445	0.13734
	0.00000	0.00001	0.00077	0.01119	0.15498
	0.00480	0.04016	0.08686	0.15828	0.35094
1	0.00482	0.04021	0.08702	0.15878	0.35175
	0.00480	0.04016	0.08694	0.15817	0.34962
	0.03297	0.11600	0.18952	0.28177	0.48979
2	0.03283	0.11533	0.18870	0.28048	0.48774
	0.03273	0.11583	0.18909	0.28136	0.48714
	0.07989	0.20238	0.29248	0.39536	0.60144
3	0.07996	0.20209	0.29285	0.39577	0.60194
	0.08021	0.20183	0.29178	0.39469	0.60173
	0.14186	0.29522	0.39575	0.50295	0.69774
4	0.14210	0.29479	0.39548	0.50249	0.69716
	0.14219	0.29453	0.39491	0.50194	0.69772
	0.21478	0.39174	0.49808	0.60425	0.78180
5	0.21395	0.39259	0.49865	0.60491	0.78312
	0.21478	0.39265	0.49850	0.60472	0.78276
	0.29836	0.49416	0.60169	0.70254	0.85679
6	0.30023	0.49498	0.60206	0.70260	0.85563
	0.30010	0.49465	0.60142	0.70191	0.85529
	0.39559	0.60149	0.70477	0.79534	0.91846
7	0.39535	0.60212	0.70508	0.79551	0.91837
	0.39711	0.60172	0.70477	0.79548	0.91840
	0.50794	0.71462	0.80739	0.88154	0.96602
8	0.50742	0.71463	0.80742	0.88150	0.96601
	0.50732	0.71523	0.80804	0.88187	0.96582
	0.64437	0.83804	0.90997	0.95778	0.99477
9	0.64576	0.83798	0.91025	0.95784	0.99471
	0.64400	0.83750	0.91000	0.95783	0.99471
	0.82922	0.98051	0.99812	0.99996	1.00000
10	0.86415	0.98556	0.99866	0.99997	1.00000
	0.81278	0.97968	0.99801	0.99995	1.00000

The R and JAGS code for the calculation of the posterior distributions is given in Appendix B.1.

3.3.3 Properties and prior choices for BHM and HBB

The next step is to expand the Bayesian models and to assume that the parameters of the prior (logit-normal or beta) are random variables by themselves and that they are described by probability distributions. This converts the models into hierarchical models.

The hierarchical structure in the beta-binomial model does not conclude in a conjugate posterior distribution with a closed analytical form anymore. Hence, numerical MCMC sampling is required to determine the posterior distribution in a HBB. Instead of using distributions for the parameters a and b, distributions for $\frac{a}{a+b}$, which is the expected value of a beta distribution, and for $a+b$, which is the effective sample size, are assumed. Hence, the position of the common distribution on the highest level is determined by the distribution of the expected value $\frac{a}{a+b}$. The amount of sharing is determined by the distribution of the ESS $a+b$. The parameter a in a beta-binomial model can be interpreted as the number of successes/responses and the parameter b as the number of failures/non-responses. The ESS $a+b$ reflects the total number of underlying observations. High values of $a+b$ reflect many observations, hence higher certainty and lower variance, which consequently results in a narrow beta distribution for the basket-individual response rates p_i, and, therefore, a higher amount of sharing among the baskets. Lower values for $a+b$ represent a lesser degree of sharing. Consequently, the amount of sharing in a HBB is determined by the distribution of $a+b$. Natural choices for the distribution of the expected value $\frac{a}{a+b}$ and the ESS are a beta distribution and a gamma distribution, respectively. The HBB for the basket-individual response rate p_i is given as

$$\begin{aligned} p_i \sim & Beta(a,b) \\ & \frac{a}{a+b} \sim Beta(\alpha,\beta) \qquad\qquad \forall i = 1,...,I, \\ & a+b \sim Gamma(\phi,\varrho) \end{aligned}$$

with hyperparameters α, β, ϕ, and ϱ.

The BHM is defined according to Berry et al. (2013) with a normal distribution for the expected value and an inverse-gamma distribution for the variance of the logit-normal distribution of response rate p_i ($\theta_i := logit(p_i)$)

$$\begin{aligned} \theta_i &\sim N(\mu, \sigma^2) \\ \mu &\sim N(\mu_\mu, \sigma_\mu^2) \qquad\qquad \forall i = 1, ..., I. \\ \sigma^2 &\sim IG(\kappa, \nu) \end{aligned}$$

The hyperparameters are $\mu_\mu, \sigma_\mu^2, \kappa$, and ν. The position of the common distribution is determined by the distribution of μ and the amount of sharing is determined by the distribution of σ^2.

The choices of the hyperparameters in both models and, consequently, the prior distributions determine how the models behave in combination with observed data. A similar starting point is required for a fair comparison of the two models. This means that the priors must ensure that the position around which the common distribution is located is similar and, more importantly, that the amount of sharing is comparable between BHM and HBB. This is a prerequisite for a comparison of BHM and HBB. The elaborated steps to achieve this similarity in the prior distributions are presented in the following.

The BHM serves as the orientation model because it is the basic model and is widely used in literature for basket trials. With reference to Berry et al. (2013), the distribution for μ is set to be centered at the fixed value $\mu_\mu := logit(0.1)$, while the variance is set to $\sigma_\mu^2 := 8^2$, which allows flexibility in the position for the common distribution. The inverse-gamma distribution is defined by the shape parameter $\kappa := 0.01$ and the rate parameter $\nu := 0.1$. This setup allows for sharing between the baskets because the posterior distributions have higher certainty than individual posterior distributions. The distribution of the expected value $\frac{a}{a+b}$ in HBB is defined in accordance to the distribution of μ from the BHM. The hyperparameters α and β for the beta distribution of $\frac{a}{a+b}$ are estimated via maximum-likelihood estimators based on $expit()$-transformed samples from the normal distribution for the position in BHM ($N(logit(0.1), 8^2)$). This estimation procedure is the one which was

elaborated in Subsection 3.3.1, and the resulting distribution for the position in HBB is $Beta(0.1377, 0.1945)$.

The transfer of the sharing property in BHM, defined by the variance $\sigma^2 \sim IG(\kappa, \nu)$, to the distribution of $a + b \sim Gamma(\phi, \varrho)$ is not straightforward. Therefore, an approximate approach is introduced. This requires a reference data scenario, which is used to calibrate the HBB with respect to the sharing behaviour of the BHM on this reference data scenario. Firstly, the reference data scenario is defined. In general, it can be any outcome of the basket trial. A guided approach is to choose the reference data scenario such that it reflects a desirable outcome of the trial. Secondly, the BHM is used to calculate the basket-wise posterior distributions. In the next step, a HBB is applied to the reference data scenario. The distribution of the expected value in the HBB is already determined as described above. The distribution of the ESS $a + b$ is varied and the resulting posterior distributions are compared to the ones from the BHM. The comparison is done using the quantiles of the distributions, especially the 2.5% and 97.5% quantiles, because the difference of these two is the width of the 95% credibility interval. The width is an indicator for the amount of sharing, with smaller widths for a high degree of sharing and vice versa. An appropriate hyperparameter choice for ϕ and ϱ is found if the 2.5% and 97.5% quantiles of HBB and BHM are similar and, most importantly, the width of the credibility intervals are of comparable length. Only similar values can be obtained for the quantiles due to the different distributional assumptions with resulting different characteristics which were shown in the previous subsections (e.g. logit-normal distribution tends towards the boundaries, beta distribution towards the center). Consequently, the width of the 95% credibility interval is the main measure to determine the hyperparameters ϕ and ϱ. In the particular scenario in this thesis, the reference data scenario is set to $r = 3$ responses among $n = 10$ observations in each of the $I = 6$ baskets. Additionally, the hyperpriors are chosen such that the credibility intervals for HBB are slightly wider, in order to ensure that HBB has at least no starting bonus. The explicit example BHM with $\mu \sim N(logit(0.1), 8^2)$ and $\sigma^2 \sim IG(0.01, 0.1)$ returns a mean 95% credibility interval of $(0.1389, 0.4980)$ with a width of 0.3591 for this reference data scenario. The mean credibility interval is calculated as the mean over the quantiles of the six baskets. A HBB with $\frac{a}{a+b} \sim Beta(0.1377, 0.1945)$ and a calibrated distribution $a + b \sim G(3, 0.16)$ results in a mean 95% credibility interval of $(0.1398, 0.5035)$ with a width of 0.3637.

The here proposed procedure for the elaboration of comparable prior distributions in BHM and HBB are applied in the next subsections, where the information sharing property in both models is compared in a simulation study.

3.3.4 Simulation scenarios and evaluation measures

The sharing property of the HBB is compared to BHM in a simulation study among four different prior scenarios and for twelve different data scenarios.

The prior scenarios are described in Table 3.8. The priors are set for BHM and then derived for HBB according to the described procedure in Subsection 3.3.3. The priors vary for the distribution of the position. An expected value of $logit(0.3)$ for the normal distribution of μ reflects a more optimistic assumption about the response rate than $logit(0.1)$. The variance is then slightly reduced to 7.8^2 because otherwise the maximum likelihood estimation for the corresponding beta distribution has numerical problems. The second variation is to assume a different reference scenario with $r = 6$ responses and $n = 20$ observations in each of the $I = 6$ baskets.

Table 3.8: *Prior scenarios for the simulations to compare BHM and HBB.*

Prior scenario		BHM		HBB	
	n	$\mu \sim$	$\sigma \sim$	$\frac{a}{a+b} \sim$	$a+b \sim$
1	10	$N(logit(0.1), 8^2)$	$IG(0.01, 0.1)$	$Beta(0.1377, 0.1945)$	$IG(3, 0.16)$
2	10	$N(logit(0.3), 7.8^2)$	$IG(0.01, 0.1)$	$Beta(0.1559, 0.1787)$	$IG(3, 0.16)$
3	20	$N(logit(0.1), 8^2)$	$IG(0.01, 0.1)$	$Beta(0.1377, 0.1945)$	$IG(3, 0.12)$
4	20	$N(logit(0.3), 7.8^2)$	$IG(0.01, 0.1)$	$Beta(0.1559, 0.1787)$	$IG(3, 0.12)$

The twelve different data scenarios are described in Table 3.9. The scenarios present a variety of possible response rates for each basket and reach from completely homogeneous to completely heterogeneous responses. For each data scenario, 10^5 outcomes are simulated. In each of the 10^5 iterations, the number of responses in every basket are sampled from a binomial distribution, assuming the respective true response rate and the number of observations given by the data and prior scenario. The simulated responses are then used to calculate the basket-individual posterior distributions with BHM and HBB. Additionally, the posterior distributions without sharing are calculated, once using a logit-normal prior and once for a beta prior. The prior distributions for the response rates in these independent calculations

are $logitN(logit(0.1), 8^2)$ and $logitN(logit(0.3), 7.8^2)$ for the logit-normal prior, while they are $Beta(0.1377, 0.1945)$ and $Beta(0.1559, 0.1787)$ for the beta prior.

Table 3.9: *Data scenarios for assumed response rates per basket in simulations to compare BHM and HBB.*

Data scenario	Basket					
	1	2	3	4	5	6
1	0.30	0.30	0.30	0.30	0.30	0.30
2	0.20	0.20	0.20	0.20	0.20	0.20
3	0.10	0.10	0.10	0.10	0.10	0.10
4	0.10	0.30	0.30	0.30	0.30	0.30
5	0.10	0.10	0.30	0.30	0.30	0.30
6	0.10	0.10	0.10	0.30	0.30	0.30
7	0.10	0.20	0.30	0.30	0.30	0.30
8	0.10	0.15	0.30	0.30	0.30	0.30
9	0.10	0.15	0.20	0.30	0.30	0.30
10	0.05	0.10	0.15	0.20	0.25	0.30
11	0.18	0.20	0.22	0.24	0.26	0.28
12	0.18	0.19	0.20	0.21	0.22	0.23

The measure to evaluate the sharing property in the simulations is the mean posterior probability to exceed the assumed true response rate which was used to simulate the data. The mean is calculated over the 10^5 simulated studies and for each basket. For data scenario 1 and basket 1, the mean posterior probability to exceed the assumed true response rate is given by

$$\frac{1}{10^5}\sum_{j=1}^{10^5} P[p_1 > 0.30|\mathcal{D}_j],$$

where p_1 is the random response rate in basket 1, 0.30 is the assumed true response rate which was used to generate the simulation data, and $\mathcal{D}_i$ contains the sampled responses from all baskets in iteration j.

This evaluation measure is calculated for each basket, in each data scenario, and in each prior scenario using the two independent calculations with logit-normal and beta prior, and the two hierarchical models, BHM and HBB. The independent calculations serve as a control, where no sharing among the baskets takes place. The changes in the mean posterior probability to exceed the true response rate between the hierarchical models and the corresponding independent calculation indicates the strength of the hierarchical sharing, because it describes

by how much the position of the posterior distribution shifts due to the sharing. The posterior probabilities for the independent beta-binomial model (BB) are calculated analytically using the conjugate property.

3.3.5 Simulation results for the sharing property of BHM and HBB

The results of the simulation study are presented in tables which contain the mean posterior probability that the response rate exceeds the assumed true response rate which was used in the simulation process. The word 'mean' is left away in the following to reduce the complexity in the description of the results. Table 3.10 and Table 3.11 show the results for the twelve data scenarios using the four different prior scenarios. The R and JAGS code for the simulation studies are given in Appendix B.2.

The first three data scenarios reflect homogeneous responses among the baskets. In the scenarios 1 and 2, the HBB shows the highest absolute posterior probability that the response rate exceeds the true response rate. It is higher compared to BHM as well as to the independent beta-binomial model. The increment in the posterior probability from independent BB to HBB is higher than the increment from the independent logit-normal prior model to BHM. Hence, HBB tends to shift the position of the posterior distribution more strongly, which means the sharing is more active. Moreover, the absolute posterior probability in BHM is smaller than the one in the independent BB, which indicates that the distributional assumptions (logit-normal or beta) for the response rate influence the form and position of the posterior probability. In data scenario 3, the sharing property of HBB is weaker compared to the other two homogeneous scenarios. This can be interpreted from the small increment of independent BB to HBB, and the small difference between HBB and BHM. The performance of HBB improves (increased difference to BHM and higher increment to independent BB) in the third data scenario when prior scenario 2 is applied, in which a shift in the position towards higher response rates is assumed. The performances of the data scenarios 1 and 2 do not change in prior scenario 2 compared to prior scenario 1. The inconsistency across the homogeneous data scenarios in the performance of HBB compared to BHM disappears when the number of observations per basket increases to $n = 20$, as is the case in prior scenarios 3 and 4. In both of these particular prior scenarios, the performances are comparable, and the change in the prior position from $logit(0.1)$ to $logit(0.3)$ only leads to a slightly higher

posterior probability throughout all four methods. Additionally, the increase in the number of observations leads to posterior probabilities closer to 0.5, which can be explained by the convergence to a normal distribution with symmetry around the expected value.

The data scenarios 4 to 6 represent two clusters with a response rate that is either 0.1 or 0.3. In all three of them, an orientation towards the overall mean response rate is observed for the hierarchical models. For true response rates below the overall mean, the posterior probability to exceed the true response rate increases compared to the independent evaluation, while for true response rates above the overall mean, the posterior probability declines. These increases and declines are stronger in HBB than those in BHM, which indicates that HBB shares information among the baskets to a higher degree. In general, the findings remain the same in the other prior scenarios. In prior scenario 2, where the position of the prior is raised, a slight increase in the posterior probabilities is observed, especially in HBB when the true response rate is 0.1. An increase in the number of observations per basket (prior scenarios 3 and 4) leads to higher posterior probabilities for baskets below the overall mean response rate and lower posterior probabilities above the overall mean response rate in all four calculation methods. The change can be explained by a decreased variance due to more observations and, therefore, an increased certainty about the position of the posterior distribution.

The scenarios 7 to 9 consist of one cluster and different individual responses in the other baskets. The individual true response rates are lower than the common response rate in the cluster. In general, the same tendency towards the overall mean response rate is observed for the hierarchical models. The HBB shows a higher activity in the shift compared to BHM, which indicates a stronger sharing of information among the baskets. In BHM and HBB, the shift towards the overall mean response rate is stronger when the true response rate is further away from the overall mean response rate. An increase of the prior position (prior scenario 2) leads in general to an increase in the posterior probabilities in all baskets for all methods. This increase is stronger in HBB when the true response rates are low (e.g. 0.1), but it disappears when the number of observations is higher, namely $n = 20$ (prior scenario 4). The general reaction to an increased number of observations per basket (prior scenarios 3 and 4) is equivalent (higher certainty) to the data scenarios before.

The scenarios 10, 11, and 12 have different response rates for each basket. The differences in the true response rates between the baskets reach from 0.05 (scenario 10) to 0.01 (scenario 12). In all scenarios, the general orientation of hierarchical models towards the overall mean response rate is observed, but now HBB does not always show the strongest tendency towards the overall mean response rate. This can be seen in scenario 10 where the absolute posterior probability to exceed the true response rate 0.05 is larger for BHM compared to HBB, and for the true response rate 0.20 it is vice versa. Also, the increment in the posterior probability from the individual analysis to the hierarchical analysis is larger for BHM compared to HBB when the true response rate is 0.05 and 0.10. In the scenarios 11 and 12, the HBB has higher absolute posterior probabilities compared to BHM, even when the true response rate is above the overall mean response rate. Additionally, the sharing strength is not persistently stronger for HBB. In data scenario 11, the decline towards the overall mean response rate is smaller for HBB compared to BHM when the true response rate is 0.22 and 0.24. In data scenario 12, this holds true for all response rates larger than the overall mean response rate.
In data scenario 10, the increased prior position in prior scenario 2 leads to a higher absolute posterior probability for the basket with response rate 0.05 under HBB compared to BHM. Moreover, the complete HBB reacts as in the previous heterogeneous data scenarios with stronger orientation towards the overall mean response rate compared to BHM. The data scenarios 11 and 12 behave analogously under both prior scenario 2 and prior scenario 1.
In data scenario 10, a higher number of observations (prior scenario 3) results in a sharing behaviour that is consistent to the previous heterogeneous data scenarios, where the HBB has a stronger tendency towards the overall mean response rate. Still, for the true response rate 0.20, the HBB has a higher absolute response rate than BHM. The data scenarios 11 and 12 behave under prior scenario 3 similarly as under prior scenario 1. Prior scenario 4 does not present any additional improvements.

The general result from the simulations is that the HBB tends to share information among the baskets to a stronger and more active degree than BHM does. This conclusion can be drawn from the absolute values and from the changes compared to independent evaluation in the posterior probability to exceed the true response rate. The absolute values describe the position of the posterior distributions, and the simulation results show that the HBB orientates the position more towards the overall mean than BHM does. This shift is quantified

by the change of the posterior probabilities in comparison to the individual assessment, where no sharing is conducted. The changes are often stronger in HBB as compared to BHM, and the change describes the sharing effect which is incorporated by the hierarchical assessment. Moreover, the simulations show that the underlying true data scenario influences the performance of BHM and HBB, while also indicating that the prior distributions and the number of observations influence the behaviour of BHM and HBB. The simulations illustrate that the HBB is an applicable tool to share information among baskets. It can therefore serve as a basic sharing tool in basket trials, equivalent to the role of BHM as the currently predominant basic sharing tool.

Table 3.10: *Mean posterior probability to exceed the true response rate under the defined data scenarios using prior scenario 1 and 2.*

	Prior scenario 1						Prior scenario 2					
1	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
indep. BB	0.4674	0.4674	0.4674	0.4674	0.4674	0.4674	0.4709	0.4709	0.4709	0.4709	0.4709	0.4709
indep. logitN	0.4481	0.4572	0.4573	0.4545	0.4545	0.4559	0.4522	0.4615	0.4616	0.4587	0.4588	0.4600
BHM	0.4564	0.4625	0.4624	0.4604	0.4607	0.4616	0.4575	0.4635	0.4635	0.4615	0.4618	0.4626
HBB	0.4774	0.4833	0.4829	0.4810	0.4816	0.4820	0.4783	0.4842	0.4838	0.4819	0.4824	0.4828
2	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
indep. BB	0.4465	0.4465	0.4465	0.4465	0.4465	0.4465	0.4505	0.4505	0.4505	0.4505	0.4505	0.4505
indep. logitN	0.4203	0.4285	0.4296	0.4282	0.4258	0.4277	0.4253	0.4335	0.4346	0.4332	0.4308	0.4327
BHM	0.4322	0.4374	0.4381	0.4370	0.4361	0.4371	0.4336	0.4387	0.4394	0.4383	0.4374	0.4385
HBB	0.4619	0.4674	0.4678	0.4666	0.4659	0.4668	0.4656	0.4709	0.4711	0.4702	0.4694	0.4699
3	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
indep. BB	0.4027	0.4027	0.4027	0.4027	0.4027	0.4027	0.4085	0.4085	0.4085	0.4085	0.4085	0.4085
indep. logitN	0.3699	0.3787	0.3782	0.3763	0.3767	0.3766	0.3767	0.3854	0.3850	0.3831	0.3836	0.3833
BHM	0.3923	0.3977	0.3966	0.3959	0.3960	0.3969	0.3943	0.3995	0.3983	0.3980	0.3980	0.3987
HBB	0.3999	0.4056	0.4050	0.4036	0.4039	0.4046	0.4174	0.4243	0.4221	0.4210	0.4217	0.4215
4	0.10	0.30	0.30	0.30	0.30	0.30	0.10	0.30	0.30	0.30	0.30	0.30
indep. BB	0.4027	0.4674	0.4674	0.4674	0.4674	0.4674	0.4085	0.4709	0.4709	0.4709	0.4709	0.4709
indep. logitN	0.3699	0.4573	0.4574	0.4544	0.4544	0.4559	0.3768	0.4615	0.4615	0.4586	0.4588	0.4601
BHM	0.7509	0.4032	0.4032	0.4007	0.4009	0.4020	0.7521	0.4043	0.4043	0.4016	0.4019	0.4030
HBB	0.8170	0.4116	0.4113	0.4091	0.4095	0.4103	0.8232	0.4123	0.4118	0.4097	0.4102	0.4108

5	0.10	0.10	0.30	0.30	0.30	0.30	0.10	0.10	0.30	0.30	0.30	0.30
indep. BB	0.4027	0.4027	0.4674	0.4674	0.4674	0.4674	0.4085	0.4085	0.4709	0.4709	0.4709	0.4709
indep. logitN	0.3700	0.3786	0.4574	0.4544	0.4545	0.4559	0.3768	0.3854	0.4616	0.4586	0.4586	0.4601
BHM	0.6750	0.6788	0.3591	0.3566	0.3567	0.3581	0.6766	0.6804	0.3599	0.3573	0.3575	0.3588
HBB	0.7316	0.7337	0.3471	0.3446	0.3452	0.3466	0.7528	0.7557	0.3466	0.3445	0.3450	0.3459
6	0.10	0.10	0.10	0.30	0.30	0.30	0.10	0.10	0.10	0.30	0.30	0.30
indep. BB	0.4027	0.4027	0.4027	0.4674	0.4674	0.4674	0.4085	0.4085	0.4085	0.4709	0.4709	0.4709
indep. logitN	0.3700	0.3787	0.3781	0.4545	0.4545	0.4558	0.3768	0.3855	0.3849	0.4586	0.4586	0.4601
BHM	0.6035	0.6077	0.6072	0.3232	0.3234	0.3249	0.6055	0.6096	0.6090	0.3236	0.3240	0.3255
HBB	0.6447	0.6484	0.6485	0.2902	0.2906	0.2918	0.6676	0.6710	0.6701	0.2884	0.2885	0.2898
7	0.10	0.20	0.30	0.30	0.30	0.30	0.10	0.20	0.30	0.30	0.30	0.30
indep. BB	0.4027	0.4465	0.4674	0.4674	0.4674	0.4674	0.4085	0.4505	0.4709	0.4709	0.4709	0.4709
indep. logitN	0.3700	0.4285	0.4573	0.4544	0.4544	0.4559	0.3768	0.4335	0.4615	0.4587	0.4587	0.4601
BHM	0.7258	0.5291	0.3778	0.3754	0.3754	0.3767	0.7271	0.5304	0.3788	0.3763	0.3764	0.3776
HBB	0.7854	0.5718	0.3786	0.3765	0.3768	0.3778	0.7960	0.5758	0.3789	0.3768	0.3771	0.3780
8	0.10	0.15	0.30	0.30	0.30	0.30	0.10	0.15	0.30	0.30	0.30	0.30
indep. BB	0.4027	0.4301	0.4674	0.4674	0.4674	0.4674	0.4085	0.4347	0.4709	0.4709	0.4709	0.4709
indep. logitN	0.3700	0.4096	0.4573	0.4544	0.4545	0.4559	0.3768	0.4152	0.4615	0.4587	0.4587	0.4601
BHM	0.7042	0.6043	0.3676	0.3650	0.3650	0.3663	0.7056	0.6057	0.3685	0.3658	0.3658	0.3671
HBB	0.7622	0.6567	0.3628	0.3606	0.3608	0.3621	0.7769	0.6659	0.3628	0.3606	0.3610	0.3620

9	0.10	0.15	0.20	0.30	0.30	0.30	0.10	0.15	0.20	0.30	0.30	0.30
indep. BB	0.4027	0.4301	0.4465	0.4674	0.4674	0.4674	0.4085	0.4347	0.4505	0.4709	0.4709	0.4709
indep. logitN	0.3700	0.4096	0.4296	0.4544	0.4545	0.4559	0.3768	0.4153	0.4346	0.4586	0.4587	0.4601
BHM	0.6810	0.5751	0.4794	0.3418	0.3420	0.3433	0.6826	0.5766	0.4806	0.3425	0.3428	0.3442
HBB	0.7307	0.6213	0.5138	0.3305	0.3310	0.3320	0.7457	0.6305	0.5184	0.3301	0.3304	0.3314
10	0.05	0.10	0.15	0.20	0.25	0.30	0.05	0.10	0.15	0.20	0.25	0.30
indep. BB	0.3464	0.4027	0.4301	0.4465	0.4582	0.4674	0.3559	0.4085	0.4347	0.4505	0.4619	0.4709
indep. logitN	0.3115	0.3787	0.4105	0.4282	0.4403	0.4559	0.3218	0.3855	0.4161	0.4331	0.4447	0.4601
BHM	0.7109	0.5683	0.4607	0.3807	0.3279	0.2975	0.7132	0.5703	0.4623	0.3818	0.3287	0.2979
HBB	0.7094	0.5914	0.4872	0.3923	0.3143	0.2551	0.7523	0.6185	0.5009	0.3984	0.3146	0.2509
11	0.18	0.20	0.22	0.24	0.26	0.28	0.18	0.20	0.22	0.24	0.26	0.28
indep. BB	0.4407	0.4465	0.4516	0.4561	0.4602	0.4639	0.4450	0.4505	0.4554	0.4598	0.4638	0.4675
indep. logitN	0.4134	0.4285	0.4369	0.4399	0.4435	0.4507	0.4186	0.4335	0.4417	0.4446	0.4480	0.4550
BHM	0.5397	0.5017	0.4638	0.4262	0.3938	0.3658	0.5412	0.5031	0.4652	0.4275	0.3949	0.3668
HBB	0.5775	0.5365	0.4939	0.4507	0.4114	0.3750	0.5815	0.5396	0.4961	0.4525	0.4125	0.3754
12	0.18	0.19	0.20	0.21	0.22	0.23	0.18	0.19	0.20	0.21	0.22	0.23
indep. BB	0.4407	0.4437	0.4465	0.4491	0.4516	0.4539	0.4450	0.4478	0.4505	0.4531	0.4554	0.4577
indep. logitN	0.4134	0.4263	0.4296	0.4303	0.4329	0.4385	0.4186	0.4314	0.4346	0.4351	0.4377	0.4431
BHM	0.4859	0.4696	0.4494	0.4282	0.4103	0.3950	0.4874	0.4709	0.4507	0.4294	0.4116	0.3962
HBB	0.5191	0.5019	0.4797	0.4566	0.4362	0.4175	0.5242	0.5061	0.4831	0.4596	0.4386	0.4193

Table 3.11: *Mean posterior probability to exceed the true response rate under the defined data scenarios using prior scenario 3 and 4.*

	Prior scenario 3						Prior scenario 4					
1	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
indep. BB	0.4773	0.4773	0.4773	0.4773	0.4773	0.4773	0.4798	0.4798	0.4798	0.4798	0.4798	0.4798
indep. logitN	0.4615	0.4709	0.4721	0.4682	0.4676	0.4690	0.4644	0.4738	0.4751	0.4712	0.4706	0.4720
BHM	0.4656	0.4722	0.4731	0.4701	0.4704	0.4709	0.4663	0.4730	0.4738	0.4708	0.4711	0.4716
HBB	0.4824	0.4888	0.4893	0.4865	0.4871	0.4872	0.4829	0.4894	0.4899	0.4871	0.4877	0.4878
2	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
indep. BB	0.4631	0.4631	0.4631	0.4631	0.4631	0.4631	0.4659	0.4659	0.4659	0.4659	0.4659	0.4659
indep. logitN	0.4418	0.4521	0.4527	0.4484	0.4483	0.4497	0.4452	0.4555	0.4561	0.4518	0.4518	0.4531
BHM	0.4477	0.4546	0.4552	0.4518	0.4525	0.4531	0.4486	0.4554	0.4560	0.4527	0.4534	0.4539
HBB	0.4750	0.4820	0.4823	0.4790	0.4799	0.4801	0.4756	0.4827	0.4829	0.4797	0.4806	0.4808
3	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
indep. BB	0.4350	0.4350	0.4350	0.4350	0.4350	0.4350	0.4389	0.4389	0.4389	0.4389	0.4389	0.4389
indep. logitN	0.4064	0.4151	0.4154	0.4130	0.4119	0.4147	0.4112	0.4198	0.4202	0.4178	0.4167	0.4195
BHM	0.4180	0.4237	0.4235	0.4221	0.4216	0.4233	0.4195	0.4249	0.4249	0.4234	0.4230	0.4247
HBB	0.4564	0.4628	0.4625	0.4607	0.4605	0.4618	0.4600	0.4664	0.4660	0.4644	0.4641	0.4654
4	0.10	0.30	0.30	0.30	0.30	0.30	0.10	0.30	0.30	0.30	0.30	0.30
indep. BB	0.4350	0.4773	0.4773	0.4773	0.4773	0.4773	0.4389	0.4798	0.4798	0.4798	0.4798	0.4798
indep. logitN	0.4064	0.4709	0.4722	0.4682	0.4677	0.4692	0.4112	0.4739	0.4751	0.4711	0.4706	0.4721
BHM	0.7733	0.4089	0.4100	0.4064	0.4068	0.4075	0.7741	0.4096	0.4106	0.4070	0.4074	0.4081
HBB	0.8485	0.4080	0.4085	0.4055	0.4059	0.4066	0.8489	0.4086	0.4091	0.4060	0.4065	0.4072

5	0.10	0.10	0.30	0.30	0.30	0.30	0.10	0.10	0.30	0.30	0.30	0.30
indep. BB	0.4350	0.4350	0.4773	0.4773	0.4773	0.4773	0.4389	0.4389	0.4798	0.4798	0.4798	0.4798
indep. logitN	0.4065	0.4151	0.4722	0.4682	0.4676	0.4691	0.4113	0.4198	0.4752	0.4711	0.4705	0.4720
BHM	0.6889	0.6933	0.3714	0.3681	0.3682	0.3691	0.6898	0.6941	0.3719	0.3687	0.3688	0.3696
HBB	0.7761	0.7795	0.3420	0.3392	0.3395	0.3404	0.7769	0.7804	0.3423	0.3395	0.3400	0.3409
6	0.10	0.10	0.10	0.30	0.30	0.30	0.10	0.10	0.10	0.30	0.30	0.30
indep. BB	0.4350	0.4350	0.4350	0.4773	0.4773	0.4773	0.4389	0.4389	0.4389	0.4798	0.4798	0.4798
indep. logitN	0.4064	0.4152	0.4154	0.4682	0.4676	0.4691	0.4112	0.4199	0.4202	0.4711	0.4706	0.4720
BHM	0.6204	0.6254	0.6258	0.3398	0.3400	0.3411	0.6215	0.6265	0.6268	0.3401	0.3405	0.3415
HBB	0.6998	0.7044	0.7046	0.2836	0.2842	0.2852	0.7018	0.7063	0.7065	0.2836	0.2843	0.2853
7	0.10	0.20	0.30	0.30	0.30	0.30	0.10	0.20	0.30	0.30	0.30	0.30
indep. BB	0.4350	0.4631	0.4773	0.4773	0.4773	0.4773	0.4389	0.4659	0.4798	0.4798	0.4798	0.4798
indep. logitN	0.4065	0.4521	0.4721	0.4682	0.4676	0.4691	0.4112	0.4556	0.4751	0.4712	0.4706	0.4721
BHM	0.7514	0.5514	0.3839	0.3804	0.3807	0.3814	0.7521	0.5522	0.3843	0.3809	0.3813	0.3820
HBB	0.8239	0.5929	0.3733	0.3703	0.3708	0.3716	0.8244	0.5935	0.3738	0.3708	0.3713	0.3721
8	0.10	0.15	0.30	0.30	0.30	0.30	0.10	0.15	0.30	0.30	0.30	0.30
indep. BB	0.4350	0.4523	0.4773	0.4773	0.4773	0.4773	0.4389	0.4555	0.4798	0.4798	0.4798	0.4798
indep. logitN	0.4065	0.4377	0.4722	0.4682	0.4676	0.4691	0.4112	0.4417	0.4751	0.4711	0.4706	0.4721
BHM	0.7264	0.6289	0.3751	0.3718	0.3720	0.3728	0.7272	0.6297	0.3757	0.3723	0.3725	0.3733
HBB	0.8038	0.6901	0.3568	0.3540	0.3544	0.3553	0.8045	0.6908	0.3573	0.3544	0.3549	0.3558

9	0.10	0.15	0.20	0.30	0.30	0.30	0.10	0.15	0.20	0.30	0.30	0.30
indep. BB	0.4350	0.4523	0.4631	0.4773	0.4773	0.4773	0.4389	0.4555	0.4659	0.4798	0.4798	0.4798
indep. logitN	0.4065	0.4378	0.4526	0.4682	0.4676	0.4691	0.4112	0.4417	0.4561	0.4711	0.4706	0.4720
BHM	0.7093	0.6025	0.4982	0.3471	0.3475	0.3484	0.7102	0.6034	0.4991	0.3476	0.3481	0.3489
HBB	0.7782	0.6560	0.5313	0.3209	0.3215	0.3225	0.7789	0.6566	0.5319	0.3215	0.3220	0.3230
10	0.05	0.10	0.15	0.20	0.25	0.30	0.05	0.10	0.15	0.20	0.25	0.30
indep. BB	0.3950	0.4350	0.4523	0.4631	0.4710	0.4773	0.4007	0.4389	0.4555	0.4659	0.4736	0.4798
indep. logitN	0.3623	0.4151	0.4381	0.4483	0.4588	0.4691	0.3692	0.4199	0.4420	0.4518	0.4619	0.4720
BHM	0.7349	0.5948	0.4805	0.3931	0.3414	0.3119	0.7361	0.5959	0.4814	0.3938	0.3420	0.3123
HBB	0.8087	0.6571	0.5199	0.4001	0.3101	0.2437	0.8144	0.6598	0.5212	0.4007	0.3100	0.2431
11	0.18	0.20	0.22	0.24	0.26	0.28	0.18	0.20	0.22	0.24	0.26	0.28
indep. BB	0.4593	0.4631	0.4665	0.4696	0.4724	0.4749	0.4622	0.4659	0.4692	0.4722	0.4749	0.4774
indep. logitN	0.4378	0.4521	0.4565	0.4574	0.4613	0.4662	0.4414	0.4555	0.4598	0.4605	0.4644	0.4692
BHM	0.5765	0.5308	0.4822	0.4347	0.3960	0.3619	0.5774	0.5317	0.4830	0.4355	0.3966	0.3627
HBB	0.6073	0.5596	0.5075	0.4551	0.4095	0.3670	0.6079	0.5603	0.5082	0.4557	0.4102	0.3675
12	0.18	0.19	0.20	0.21	0.22	0.23	0.18	0.19	0.20	0.21	0.22	0.23
indep. BB	0.4593	0.4612	0.4631	0.4649	0.4665	0.4681	0.4622	0.4641	0.4659	0.4676	0.4692	0.4708
indep. logitN	0.4377	0.4487	0.4527	0.4521	0.4528	0.4564	0.4414	0.4522	0.4562	0.4555	0.4562	0.4596
BHM	0.5149	0.4932	0.4683	0.4417	0.4181	0.3976	0.5158	0.4941	0.4692	0.4425	0.4189	0.3984
HBB	0.5436	0.5214	0.4954	0.4672	0.4416	0.4179	0.5443	0.5222	0.4962	0.4679	0.4423	0.4185

Chapter 4

Discussion

In this chapter the results of this thesis are discussed and the structure in this chapter is according to the order in the results Chapter 3. In each section the corresponding results including the contribution to research are discussed as well as limitations and directions for further research. The discussion chapter ends with a conclusion of the results from this thesis.

4.1 Categorization and modular construction of basket trials

Parts of this Section 4.1 are already published in the article *Categories, components, and techniques in a modular construction of basket trials for application and further research* by Pohl et al. (2021). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

4.1.1 Discussion and contribution to research

The dynamic research in the past years on basket trials has led to many designs and techniques, and in general to a diffuse landscape of options for basket trials. To disentangle this situation, a categorization and a modular approach for basket trials was developed. The categorization and the modular approach offer a systematic approach to basket trial designs. The categorization of designs based on the applied statistical technique and on the purpose of the trial allows a quick and informative classification of a design. It serves as a starting

point to get a first impression of a design and whether it potentially fits to an underlying medical research question.

The modular construction has its focus on a systematic approach to a newly designed basket trial. It consists of four components which can be individually arranged. The components are then filled with tools from different already known basket trial designs or with newly created tools. A common notation was introduced among the statistical tools in this work and in the corresponding publication in Pohl et al. (2021). The understanding of the tools is required to create a new basket trial in a modular manner. The consistent notation is therefore an important cornerstone to understand the ideas of the tools, especially those for sharing of information. The technical ideas of available tools for each component were presented in Section 3.1 and serve as catalogue to look up the tools. The sharing is the key component of a basket trial and the available tools can be complex. The modular approach reveals different underlying techniques. The frequentist *pool all or nothing* approaches must be considered critically, since only two extremes of sharing (all or nothing) are allowed. The Bayesian tools allow intermediate amounts of sharing, especially the variance-driven BHMs present an intuitive technique due to their mechanism in which a higher variance corresponds to lower sharing and vice versa. The amount of sharing can be tuned by the parameter choice. The flexibility of the Bayesian tools also comes to light when information between only a subset of baskets is shared. Hobbs and Landin (2018), as well as Psioda et al. (2019) introduced these multisource exchangeability models based on the theoretical work of Kaizer et al. (2017) and many other tools (e.g. Fujikawa et al., 2020; Jin et al., 2020b; Chen and Lee, 2020) addressed the same idea with different techniques. The choice of the sharing tool is an important aspect for the acceptance of the trial design among all involved researchers. Tools which share information without a transformation of the response rate are for the acceptance of advantage, because they can use the conjugate property of a beta-binomial combination of prior distribution and observed data. In the latter case, the observations can be added in weighted form to the parameters of the posterior distribution, which makes the sharing process more understandable for non-statisticians. The interim futility and efficacy component are optional, but can help to save resources. They must be chosen in the context of the underlying trial. For example, in an exploratory phase II trial, the goal is to detect

promising baskets, but it is not of interest to already declare early efficacy, which might even not be possible due to a rather small number of patients.

The categorization and the modular construction facilitate access to basket trial designs for statisticians and interested medical staff with the intention to accelerate the practical application of basket trials, to facilitate communication about them, and to motivate further research due to increased visibility of the current landscape of available tools and their connections.

4.1.2 Limitations and directions for further research

The categorization is limited to the so far available trial designs and therefore reflects the current state of research. It can be assumed that the number of metrics to efficiently categorise basket trial designs might increase in the future, and new values within the metrics might be added. Potentially new aspects could be phase I dose finding trials or seamless trial designs. The current categorization is not a static rule and should evolve as dynamically as the whole research field does. The categorization also revealed that there are only few basket trial designs available for phase III and when they are, frequentist techniques are used. This reflects the current expectations of regulatory agencies for the approval of new treatments. However, the quickly evolving field of personalized medicine could change that and when this is the case, efficient basket trial designs should be available. This includes sharing tools which are more sophisticated than the frequentist *pool all or nothing* approaches. Bayesian techniques are of advantage in that case, hence further research for Bayesian phase III designs might be of interest. This also includes sharing techniques for endpoints that reflect the patients' benefit more precisely, e.g. in form of a time-to-event endpoint. Such an endpoint will additionally increase the complexity in the interim analyses, which might constitute another field of future research. Additional research efforts can be invested in seamless designs, e.g a combined investigation of dose levels and first evidence of efficacy in a phase I/II design, for which the recent publication by Lin et al. (2021) could be a starting point. Also, seamless phase II/III designs where promising baskets from phase II directly switch to phase III can be subject to further research.

The modular framework led to subcategories for the Bayesian sharing techniques (mean-drive, variance-driven, hypothesis-driven, non-transformed) and each of them can be a field

for further research with the intention to create efficient sharing techniques with reduced complexity. A high complexity of the tools turns basket trials into a black box where the stakeholders of the trial barely understand what is happening with the data. This is a current limitation for the application of innovative sharing tools in real basket trials. Moreover, the choice for a sharing tool must be justified by advantageous characteristics compared to other tools. However, the innovative sharing tools were mostly only compared to the basic tools. Hence, a broad comparison of all relevant sharing tools to each other is necessary, and is an important aspect to transfer the theoretical tools into applications in real trials.

Another limitation is that programming code is not easily accessible for every tool. Even if code is available, it is stored in different places. Hence, future research should focus on the creation of a package for the open-source software R which should contain the known tools such that a modular basket trial can easily be constructed. With such a package, design characteristics of a modular basket trial can be investigated in simulations and also comparisons among different designs would be possible.

Generally, the field of further research in basket trials is wide and covers the optimal arrangement of components, the improvement and modification of existing tools for the components, and the development of new innovative methods which improve the performance of a basket trial. This comprises analytical investigations as well as numerical ones using extensive simulation studies.

4.2 Relationships among decision tools in basket trials

4.2.1 Discussion and contribution to research

The investigation of the Bayesian and frequentist decision rules showed that the adaption of the threshold leads to congruence in the decisions based on the one-sided binomial test and on the Bayesian posterior distribution of a beta-binomial model with uniform prior. The threshold of the posterior probability must be adapted dynamically based on n and p_0, which means that local control is possible. A global choice for the threshold is only given for a beta prior with $a = 0, b = 1$, because then, the p-value and the posterior probability are the same for every n, r and p_0. Local control is sufficient in practical scenarios of basket trials, where,

e.g. the number of patients per basket or in total is prespecified to a certain range of values. These are tuning parameters of a basket trial that need to be defined in the context of the trial objective and based on pragmatic restrictions.

The required adaption of the threshold is based on the general difference of the p-value and the posterior probability. The analytically elaborated differences specify this difference more clearly and they are presented in form of a correction factor δ and the absolute difference κ. Both quantify the general difference when a uniform prior is applied. They enable to directly convert $\mathcal{B}$ into $\mathcal{P}$ and vice versa. The advantage of the correction factor is that a relative difference is shown according to the basis, which is the p-value in this case. However, for small values the correction factor can be very large while the absolute difference is small, therefore both measures are used to evaluate the difference. The quantification of the difference allows $\mathcal{B}$ and $\mathcal{P}$ to be displayed as functions of each other. Especially the posterior probability as a function of a p-value allows to convert a frequentist design into a basket trial with Bayesian decision rules. In the other direction, the posterior probability can be converted into a p-value such that a comparison to a constant threshold is possible irrespective of the observed data. The conversion of the frequentist example of Cunanan et al. (2017b) into a design with Bayesian tools shows the practical purpose of the (absolute and relative) difference in the setting of a basket trial. Finally, the frequentist and the Bayesian tool can simply be considered as two different tools to display the observed data and to take decisions.

The posterior probability could also be calculated directly based on the available data, however, the multiplicative or additive adaption of the p-value shows by how much the two tools differ. This also shows that the naive approach to use $1 - \alpha$ as the threshold for efficacy assessment (e.g. in Psioda et al., 2019) is misleading and results in overoptimism of Bayesian decisions which consequently supports the negative impression that Bayesian statistics does not take enough care of error control. The use of $1 - \alpha$ can only approximately achieve a false-positive rate of α because the absolute difference declines to 0 for $n \to \infty$. However, for even a moderately high number of observations in a basket (e.g. $n = 30$, $p_0 = 0.30$), the absolute difference δ is substantially far away from 0, which means the same error control can then not be assured. The differences are high in response regions in which discriminatory decisions are made. Especially in these regions, it is important to mind the difference.

The analytical relationship was elaborated with respect to efficacy ($p > p_0$) but can be converted to the assessment of futility by changing the reference value and/or the threshold τ. Hence, the differences in frequentist and Bayesian tools and the congruence of decisions also hold true in the assessment of futility.

In Cunanan et al. (2017b) the interim futility assessment requires an arbitrary minimum number of responses to continue the trial (or a specific basket). This tool is a naive approach and easily understood, however, it does not directly correspond to a certain statistical metric (e.g. p-value, posterior probability) that can justify the threshold choices. In the conversion of the practical example, this requirement is transformed into a posterior probability via the corresponding p-value from a one-sided binomial test. In the heterogeneous path with only a few patients per basket, this results in rather wide intervals from which the threshold values can be chosen. Choosing a value close to the borders might appear surprising when there is unawareness about the interval range. Moreover, in the case of unequal accrual, the required one response among 3 to 10 patients reflects quite different response scenarios in the heterogeneous path, and the plausibility of the minimum and maximum rule for the number of patients per basket must be considered carefully in the context of the trial purpose (e.g. imbalance due to rare disease property) and in accordance to available resources (e.g. possibility to wait for equal numbers in each basket). For the practical conversion under unequal accrual, the interim requirement imposes a minor restriction because there is no threshold value τ_F that can cover the wide spectrum for $r = 1$ and $n \in \{3, ..., 10\}$. However, a unique solution is available if $n = 3$ or $n = 10$ is removed and the minimum and maximum rules can be defined accordingly. Apart from that, the example shows that the conversion works well in general, and threshold values can be determined for each node of the trial such that congruence in the decisions is guaranteed, irrespective of equal or unequal numbers per basket.

The conversion of the example shows that Bayesian tools do not change the error rates, because the new tools will exactly take the same decisions. This also means that the T1E rates will also be controlled when Bayesian tools are applied. Bayesian statistics is not relieved from T1E control, they take decisions as well and need to inform how well they do, since error control is a fundamental characteristic of a decision tool. In frequentist statistics, the T1E control is required and due to the direct link to the Bayesian tools, it must be considered

obligatory there too. Hence, in practice one can plan a trial in a frequentist way and then convert it later into a basket trial with Bayesian tools. Moreover, the conversion allows a more intuitive interpretation of the final analysis, because the posterior probability presents what is actually asked: "How likely is the alternative hypothesis?" whereas the p-value can only answer whether the null hypothesis is rejected. Also, the conversion opens the door for other tools (e.g. posterior predictive probability) that could be applied for decision making, thus potentially leading to more understandable and interpretative decisions.

Apart from the naive binomial test and the posterior probability, there are other tools which were proposed for interim futility assessment. For these tools, connections among each other were elaborated. The existence of these links show that the futility tools are different ways to display the same data. The characteristics of the tools are tuned via the tool-specific parameters (reference values, thresholds). The predictive posterior probability (PPP) and the conditional power are interesting tools because they project the current information to the final analysis and quantify how likely the desired result will be achieved. The advantage of the Bayesian tools is that the posterior distribution already contains the shared information. This holds also true for the predictive posterior probability because it is a function of the posterior distribution and the number of additional observations. In the case of a beta-binomial model, a closed analytical form for the PPP is available, and for other cases simulations are feasible. The advantage of PPP over the posterior probability is that the decision threshold can be kept constant at each interim assessment because the final analysis is the target. The threshold for the posterior probability must be adapted according to the available data in order to account for the underlying uncertainties. The choice of the interim futility tool should be made in the context of the trial. A 'minimum r responses' rule can have its justification in trials where simplicity and pragmatism are important and also when the first interim is early in the trial. Kopp-Schneider et al. (2019) for example converted the posterior probability in interim assessments into a minimum number of responses and argue that this can facilitate the communication with non-statisticians. On the other side conditional power and PPP offer more informed decisions but also require a certain amount of information, including the fixed number of observations at the end of the trial.

4.2.2 Limitations and directions for further research

The quantified differences refer to the beta-binomial model with prior $a = 0, b = 1$ and $a = b = 1$. The prior $beta(0, 1)$ is hardly used in practice whereas the uniform prior $beta(1, 1)$ is applied in proposed basket trial designs (cf. Fujikawa et al. (2020)). An interesting prior choice is the Jeffrey's prior $beta(0.5, 0.5)$ for the beta-binomial model. It is considered as non-informative because it is invariant for transformations of the parameter (Robert et al., 2009; Yang and Berger, 1997) and has a reduced effective sample size (ESS) in the beta-binomial model compared to the uniform prior. This is of advantage when the number of observations is small. The characteristics of the beta-binomial model in basket trials under different priors, including the Jeffrey's prior, is of interest in further research. This includes also the behaviour of δ and κ under different priors. In general, the interpretation of the posterior must always take the prior distribution into account.

The example of Cunanan et al. (2017b) reveals that the conversion has its practical limitations for small number of observations n because then, the p-value and the posterior probability are difficult to interpret due to the high variability and the uncertainty of the data. Additionally, for small n, the prior can have substantial impact on the posterior distribution, e.g. uniform $beta(1, 1)$ prior and a posterior probability of $\mathrm{P}[p > 0.15|n = 3, r = 0] = 0.522$. The prior might out-weight the observed data and influence the decision. The naive decision rule to not stop for futility if at least a certain number of observations has been observed therefore has its merits when only few observations are available. In that case, the conversion must be considered carefully.

The conversion applies to stand-alone decisions where the underlying data is the same for the frequentist binomial test and the Bayesian posterior distribution. This is the case in basket trials where an *all or nothing* approach for the sharing is used. Most Bayesian sharing tools (e.g. hierarchical models) partially share information between baskets and consequently the amount of information in the posterior distribution is larger than in the statistical test and, therefore, a direct analytic relationship cannot be derived anymore. The basket-individual posterior distribution then depends on the results in that basket, on the results in the other baskets, and on the sharing technique.

The conversion and also the application in the example showed that the Bayesian tools are not free from the discreteness of the binomial data. The minimum response tables (e.g. Table 3.4) underlined that, since for the decision making, the discrete counts (r responses, n observations) are transformed to a value (p-value, posterior probability) on a continuous scale, which is subsequently used to justify the decisions. Consequently, the p-value and the posterior probability can be considered as just two different ways to display the observed data.

The map of connections among tools for interim (efficacy, futility) assessments is limited by the already discussed aspect of different underlying data which is induced by partial sharing. This restricts the connections between the frequentist and the Bayesian tools. A limitation of the conditional power and of the predictive posterior probability is the number of additional observations until the end of the trial. This number must be fixed and prespecified and might contradict the dynamical appearance of basket trials. The PPP accounts for the sharing of information up to the point where the interim analysis takes place but not when the final analysis is calculated under the different scenarios. Therefore, future research could investigate how the sharing of information could be incorporated and potentially help to take a more informed interim decision.

4.3 Hierarchical beta-binomial model in comparison to BHM for information sharing in basket trials

4.3.1 Discussion and contribution to research

The BHM sharing tool by Berry et al. (2013) calculates the posterior distributions on a logit-transformed scale. Nevertheless, the authors conduct the inference on the probability scale and justify this with better clinical interpretation, however, they do not argue why the logit-transformation of the response rate bears advantages in the calculations. The sharing tools which directly evolved from the BHM neither give an explanation. Hence, a vital investigation whether a hierarchical model directly on the probability scale is a suitable alternative to a logit-transformed model (BHM) in basket trials was conducted. The motivation to investigate this stems from the non-transformed model proposed by Fujikawa et al. (2020) and from the

idea that a non-transformed distribution is easier to be understood. A better comprehension of the underlying statistical tool makes it more appealing for the use in applied trials because more stakeholders of the trial understand what kind of calculations are conducted.

The independent investigation of the assumed distributions for the response rate in BHM and HBB show that the logit-normal distribution and a beta distribution are in general different distributions. The counterexample was derived via the transformation of the densities. The logit-transformation of a beta density function does not result in a normal distribution, hence the two distributions are in general different, even though they asymptotically converge towards each other. However, the number of observations per basket is usually small in basket trials and, therefore, asymptotic convergence of logit-normal and beta distribution is not reached in practice. In application it must be clear that two different distributions for the response rate p are assumed. Nevertheless, a procedure to approach a given logit-normal distribution with a beta distribution, and vice versa, was elaborated. The procedure uses MLEs to approach the given distribution (e.g. logit-normal) with the other distribution (e.g. beta). The similarity of two prior distributions is a prerequisite if a comparison of them in a Bayesian analysis is planned. This can be achieved with the proposed approach and allows for at least a similar starting point in a Bayesian comparison.

A vague prior was chosen for the comparison of the Bayesian behaviour of a logit-normal and a beta prior together with binomial data. The intention behind this choice is to let the observed data carve the posterior distribution. The results reveal that the different distribution structures pertain, and the two prior distributions result with the same binomial observations in different posterior distributions.

The HBB has the advantage of interpretability of the common beta distribution via the expected value and the effective sample size. Both of them represent concrete values that describe the common beta distribution from which the basket individual distributions of the response rates p_i are drawn. Especially the effective sample size is of advantage, because it is an equivalent to an underlying observation count. Hence, their distributions, prior and posterior, have an intuitive interpretation. On the other hand, in BHM neither the parameter μ nor the variance σ^2 have an interpretation that directly refers to intuitive measures. This

is because they refer to the normal distribution of the logit-transformed response rate and, hence, a straightforward interpretation on the probability scale is not available.

The focus in the comparison of HBB and BHM lies in the sharing characteristic. A fair comparison of the two tools under different data scenarios requires hyperparameter choices such that both tools have similar sharing characteristics. The calibration procedure for the two tools elaborated in this thesis, is supposed to ensure this. The idea is to define one reference data scenario under which both tools share information to a similar extent. The reference data scenario has the same response in each basket because this reflects the implicit exchangeability assumption of hierarchical models. It was a deliberate choice to allow the HBB to perform a bit worse than the BHM under the reference data scenario, because the HBB is challenging the widely used BHM and therefore it should not have a starting bonus. The calibration based on the sharing strength is different from the calibration on the trial characteristics, mostly the false positive rate under a null scenario. The latter is used when the complete trial design is investigated (cf. Chu and Yuan, 2018a; Chen and Lee, 2019). A calibration based on the complete trial was also used in Freidlin and Korn (2013), where the BHM was compared to a hierarchical beta-binomial model. However, the authors investigated the final analysis of the basket trial and its false positive rate. Additionally, they assumed truncated uniform distributions for a and b. Both aspects are different from the investigation in this thesis.

The simulation scenarios were chosen to reflect a wide variety of possible response rates among the baskets. The two anchor response rates 0.1 and 0.3 represent a non-promising and a promising treatment response and the data scenarios are constructed around them. The number of baskets was set to $I = 6$ and is similar to the simulation scenarios in e.g. Jin et al. (2020a) or Asano and Hirakawa (2020). The mean posterior probability to exceed the underlying true response rate was introduced to measure the sharing property of the tools. It directly refers to the true response rate and gives an impression about the position of the posterior distribution. The advantage of a comparison against the true underlying response rate is that the posterior probability to exceed it lies in a range around 0.5, whereas a comparison against a value which is far away from the underlying true response rate results in values close to 0 or 1. Moreover, in the latter cases, an evaluation of the sharing property

is barely possible because the changes in the position of the posterior distribution induced by sharing are small. This is due to the fact that most of the probability mass of the posterior distribution already lies on the left or on the right of the reference value when no sharing is conducted. In that case, the sharing property cannot be assessed. However, this is not the case when the true response rate is the reference value and, therefore, differences in the sharing tools can be detected more easily. The comparison to the individual posterior probability shows by how much the posterior distribution changes due to the sharing.

The simulation results show that in HBB the orientation towards the overall mean and, therefore, the sharing is stronger than it is in BHM. This is of advantage when all baskets are homogeneous, but can conclude in higher basket-individual false-positive rates when at least one basket is different from the others. This characteristic is what BHM is criticised for when used as a standalone sharing tool (e.g. Freidlin and Korn, 2013; Chu and Yuan, 2018a). The simulations also showed that the underlying assumption about the distribution of the response rate persists, and these distributional characteristics of a logit-normal and a beta distribution are also detected in the hierarchical models.
The results prove the feasibility of HBB to be applied as a basic sharing tool. Its characteristics are a bit different to BHM, but still they encourage a further investigation, especially when the interpretational advantage of HBB is kept in mind.

4.3.2 Limitations and directions for further research

Further research, especially for basket trial designs, must keep in mind that a normal distribution of the logit-transformed response rate is a different distributional assumption than a beta distribution for the non-transformed response rate. The logit-normal and the beta distribution can be approximated by each other, but still the differences in the distributions can be substantial, and the approximation towards each other cannot eliminate the structural differences of two distribution.

The numerical problems limit the calculation of the posterior distribution for the beta-binomial model with MCMC sampling via JAGS. Firstly, unintended stops of the MCMC sampling terminate the code early and without a result. This can be avoided with a truncated beta distribution in the JAGS code. Secondly, there are numerical instabilities in the

calculation of posterior probabilities when responses are not at all or exclusively observed. The instabilities can be overcome with higher sampling numbers, however, this also increases the calculation time, which is already higher than for the logit-normal prior. The instabilities are reduced when the number of observations increases, e.g. for $n = 20$. Further research can concentrate on a faster code for the MCMC sampling, e.g. with a different software. Albeit all those numerical drawbacks, the posterior probability can in practice be calculated analytically due to the conjugate property of a beta-binomial model.

Further methodological research of the HBB can address the distribution of the expected sample size of the common distribution. Interesting aspects could be HBBs where the gamma distribution of the ESS is for example restricted to values which are smaller than the total number of observations in the basket trial. Moreover, the posterior distribution of the ESS could be investigated as a potential measure of how much information was shared between the baskets.

A limitation of the proposed calibration procedure is that only one reference scenario with fixed numbers of responses in each of the baskets is used. Future research could concentrate on a calibration procedure based on simulations with given response rates per basket. Apart from the assumed data scenario, the assessment of similarity of the posterior distributions for the calibration procedure can be subject to additional research, e.g. calibration via minimisation of distance measures for distributions. In general, the calibration of sharing tools is an interesting field for future research because equal starting points are an essential element in the comparison of the tools and the latter is needed to enable a thorough comparison of existing and future basket trial designs.

A limitation and consequently a potential subject for further research are the simulation scenarios. Only two different number of patients per basket were simulated, also the number of patients per basket were equal. Hence, in future simulations, imbalanced number of patients per basket could be investigated. Moreover, other response rates can be considered, e.g. rates above 0.3. However, the purpose in this thesis was to investigate whether a HBB is in general a feasible basic tool to share information in a basket trial. Hence, it suffices to use it in data scenarios that can potentially be obtained in applied basket trials, and it can be argued that not every possible data scenario must be covered to come to a conclusion.

The simulations showed that HBB is a feasible basic sharing tool, but should not solely be applied, since it has the same disadvantages as BHM. Hence, future research can concentrate on BHM-based sharing tools where the BHM is replaced by a HBB. Due to the advantage in simple application and interpretation of the parameters, it suffices if the HBB modified designs do not perform inferior to the existing BHM-based designs. Also, new trial designs using HBB as the starting point might be developed. A drawback of HBB compared to BHM is the higher numeric computation time. This is no problem in the analysis of an applied basket trial, but might be an issue when simulations of the trial characteristics are performed.

4.4 Conclusion

In conclusion, basket trial designs and its tools are a diverse and dynamic field of research. The elaborated systematic approach towards basket trials via categories and modular components clarified the current status of basket trial designs and also pointed out the methodological complexities in the designs and tools. Unnecessary complexity is an obstacle for the implementation of these tools in applied trials. Therefore, the interim tools and the sharing tools were investigated with the intention to reveal unexpected connections and to minimize complexity. The interim decision rules using frequentist and Bayesian techniques turn out to be different ways to display the observed data, and it was shown that, under appropriate assumptions, the rules can be tuned to come to the same conclusions. For the sharing tools, it was shown that a hierarchical beta-binomial model is a feasible alternative basic model compared to the mainly used hierarchical logit-normal model. With the HBB, the complexity in the sharing tools can be reduced because the response rate is directly modeled without the logit-transformation. This makes an information-guided definition of the hyperparameters in the hierarchical model easier and makes the sharing tool more understandable for all involved researchers. Moreover, a calibration procedure for an equal sharing among different hierarchical models was proposed, which can be helpful in future research when different sharing tools and designs are compared to each other.

Chapter 5

Summary

5.1 Summary

The scientific advances in medical research in the last two decades have shifted the focus to a personalized treatment approach. An immanent consequence is the need for clinical trials which cover these new treatment approaches. A group of clinical trial designs which account for this are gathered under the generic term master protocols. The basket trial design has evolved as the most prominent master protocol design and investigates one treatment in several different diseases. The joint investigation is justified by a common characteristic, such as a genetic aberration, which is prevalent in all of the diseases and which is used as an effect pathway by the investigated treatment. Basket trial designs have been a virulent field of research with respect to the statistical tools and characteristics of such trials. This has led to an unclear situation in literature and an increasing level of complexity in the statistical tools which are proposed for the use throughout a basket trial. However, a practical application as well as an increased interest in basket trials is rather hampered if the complexity is increased and if no access point to the topic is available. Hence, the aim of this thesis was to introduce a systematic approach to basket trial designs and to investigate the statistical tools in order to facilitate, connect, and improve statistical tools, all with the intention to make the complete basket trial setting more accessible, understandable, and applicable from a statistical perspective. The here elaborated systematic approach towards basket trial designs consists of

two aspects, first a categorization of the trial designs and second a modular construction kit for basket trials. The categorization of basket trials is based on the purpose of the trial and on the statistical techniques that are applied. The modular construction separates a basket trial into four different components and presents available statistical tools for the components in a common notation. It moreover elaborates the methodological connections among the sharing tools and shows that they use different techniques. However, even though their complexity varies strongly, the tools are connected with each other or can be the same, even if they were proposed in different ways in different publications. The modular construction kit additionally serves as a catalogue to look up the available statistical tools when a basket trial is planned. The decision tools in basket trials were investigated with a focus on the difference in the statistical methodologies, namely between the frequentist one-sided binomial test and the Bayesian decision based on the posterior distribution from a beta-binomial model. It was shown that the decision tools can be tuned such that the same decisions are made. The difference between the frequentist p-value and the Bayesian posterior probability under a uniform prior was quantified analytically and it was shown by how much the two decision measures deviate from each other. With the elaborated difference, the p-value and the posterior probability can be given as functions of each other and therefore can be used interchangeably. The practical feasibility of that relationship for basket trials was shown with the conversion of the decision tools in a frequentist design into Bayesian decisions. Additionally, the connections between the other decision tools from the construction kit were investigated. The construction kit showed that the hierarchical model with normally distributed, logit-transformed response rate is the base for the majority of the sharing tools. In this thesis, a detailed investigation of a hierarchical model directly relying on the beta distributed, non-transformed response rate was conducted with respect to its feasibility as a basic sharing tool in basket trials. It was shown that the non-transformed model shares information to a slightly stronger degree, that the different underlying distributional assumption for the response rate persists, and that, in general, it is a feasible sharing tool which does have advantages in the interpretation of the hyperparameters. Therefore, its use in basket trial designs should be further investigated in future research. To conclude, this thesis provides a thorough investigation of basket trial designs, it starts with the elaboration of a systematic approach to them and continues with the investigation of the particular components and their statistical tools.

5.2 Zusammenfassung

Die wissenschaftlichen Fortschritte in der medizinischen Forschung in den vergangenen beiden Dekaden haben den Fokus zu einem personalisierten Behandlungsansatz verschoben. Eine immanente Konsequenz daraus ist die Notwendigkeit klinischer Studien, die diese neuen Behandlungsansätze abdecken. Eine Gruppe von Studiendesigns, die dies berücksichtigen, sind unter dem generischen Term Master Protokolle zusammengefasst. Die Basketstudie hat sich als das bisher bedeutendste Master Protokoll Design herausgestellt und untersucht eine Behandlung in mehreren unterschiedlichen Erkrankungen. Die gemeinsame Untersuchung wird über eine geteilte Charakteristik, zum Beispiel eine genetische Anomalie, begründet, welche in allen Erkrankungen vorliegt und von der untersuchten Behandlung als Wirkungsmechanismus genutzt wird. Basketstudien waren und sind in Bezug auf die statistischen Werkzeuge und auf die Eigenschaften der Studiendesigns ein dynamisches Forschungsfeld. Dies hat zu einer unübersichtlichen Literaturlage und einer zunehmenden Komplexität in den statistischen Werkzeugen geführt, die für die Anwendung in einer Basketstudie vorgeschlagen wurden. Eine praktische Anwendung, wie auch ein erhöhtes Interesse an Basketstudien, ist jedoch beeinträchtigt, wenn die Komplexität konstant erhöht wird und zudem kein Zugangspunkt zu diesem Thema gegeben ist. Daher war das Ziel dieser Dissertation, einen systematischen Zugang zu Basketstudien einzuführen und die statistischen Werkzeuge zu untersuchen, um diese zu vereinfachen, miteinander in Verbindung zu bringen und zu verbessern und all dies mit der Intention das Studiendesign aus einer statistischen Perspektive zugänglicher, verständlicher und einfacher anwendbar zu machen. Der hier erarbeitete systematische Zugang zu Basketsstudien besteht aus zwei Aspekten, erstens, der Kategorisierung der Studiendesigns und zweitens dem modularen Baukasten für Basketstudien. Die Kategorisierung basiert auf dem Studienzweck und auf den statistischen Techniken, die angewendet werden. Der modulare Baukasten unterteilt eine Basketstudie in vier Komponenten und stellt die gegebenen statistischen Werkzeuge für die einzelnen Komponenten in einer einheitlichen Notation dar. Zudem wurden methodische Verbindungen zwischen den Werkzeugen zum Teilen von Informationen zwischen den einzelnen Baskets erarbeitet und es zeigt sich, dass dafür unterschiedliche Techniken verwendet werden. Obwohl deren Komplexität stark variiert, sind diese Werkzeuge miteinander verbunden oder können sogar gleich sein, auch wenn

sie auf verschiedene Weisen und in verschiedenen Publikationen vorgeschlagen wurden. Der modulare Baukasten dient zusätzlich als ein Katalog zum Nachschlagen vorhandener statistischer Werkzeuge, insbesondere in der Planungsphase einer Basketstudie. Die Entscheidungsregeln in Basketstudien wurden mit dem Fokus auf den Unterschied in den statistischen Methoden untersucht, speziell bezüglich des frequentistischen, einseitigen Binomialtests und der Bayesianischen Entscheidung basierend auf der Posterior-Verteilung eines Beta-Binomial Modells. Es wurde gezeigt, dass die Entscheidungsregeln so eingestellt werden können, dass dieselben Entscheidungen getroffen werden. Der Unterschied zwischen dem frequentistischen p-Wert und der Bayesianischen Posterior-Wahrscheinlichkeit gegeben einem gleichverteilten Prior wurde analytisch quantifiziert und es wurde gezeigt, wie die beiden Entscheidungsmaße voneinander abweichen. Mit der erarbeiteten Differenz können der p-Wert und die Posterior Wahrscheinlichkeit als Funktionen voneinander dargestellt werden und daher zueinander austauschbar verwendet werden. Die praktische Anwendbarkeit dieser Verbindung wurde mit der Transformation der Entscheidungswerkzeuge in einem frequentistischen Studiendesign in Bayesianische gezeigt. Zudem wurden die Verbindungen zwischen den weiteren Entscheidungsregeln aus dem Baukasten untersucht. Der modulare Baukasten hat gezeigt, dass das hierarchische Model mit normal-verteilter, logit-transformierter Responserate die Basis für die Mehrheit der Werkzeuge zum Teilen von Informationen ist. In dieser Dissertation wurde eine detaillierte Untersuchung eines hierarchischen Modells direkt auf einer beta-verteilten, nicht-transformierten Responserate durchgeführt, mit Bezug auf dessen Mach- und Umsetzbarkeit als ein Basiswerkzeug zum Teilen von Informationen in Basketstudien. Es wurde gezeigt, dass das nicht-transformierte Modell die Informationen etwas stärker teilt, dass die unterschiedlichen Verteilungsannahmen für die Responserate bestehen bleiben und dass es im Allgemeinen ein anwendbares Werkzeug zum Teilen der Informationen ist, welches Vorteile in der Interpretation der Hyperparameter hat. Daher sollte seine Anwendung in Basketstudien als Teil zukünftiger Forschung weitergehend untersucht werden. Zusammenfassend bietet diese Dissertation eine eingehende Untersuchung von Basketstudien, welche mit der Erarbeitung des systematischen Zugangs beginnt und mit der Untersuchung der einzelnen Komponenten und den zugehörigen statistischen Werkzeugen fortgeführt wurde.

Bibliography

Agresti, A. (2007).

An Introduction to categorical data analysis.

John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition.

Arens, T., Hettlich, F., Karpfinger, C., Kockelkorn, U., Lichtenegger, K., and Stachel, H. (2009).

Mathematik.

Spektrum Akademischer Verlag, Heidelberg.

Asano, J. and Hirakawa, A. (2020).

A Bayesian basket trial design accounting for uncertainties of homogeneity and heterogeneity of treatment effect among subpopulations.

Pharmaceutical Statistics, 19(6):975–1000.

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. J. (2009).

I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy.

Clinical Pharmacology & Therapeutics, 86(1):97–100.

Bayes, T. and Price (1763).

An essay towards solving a problem in the doctrine of chances.

Philosophical Transactions of the Royal Society of London, 53:370–418.

Berry, D. A. (2015).

The brave new world of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research.
Molecular Oncology, 9(5):951–959.

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.

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.

Chen, N. and Lee, J. J. (2019).
Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes.
Biometrical Journal, 61(5):1219–1231.

Chen, N. and Lee, J. J. (2020).
Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints.
Statistical Methods in Medical Research, 29(9):2717–2732.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011).
Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.
Texts in Statistical Science. Chapman & Hall/CRC, 1st edition.

Chu, Y. and Yuan, Y. (2018a).
A Bayesian basket trial design using a calibrated Bayesian hierarchical model.
Clinical Trials, 15(2):149–158.

Chu, Y. and Yuan, Y. (2018b).
BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity.

Journal of the Royal Statistical Society: Series C (Applied Statistics), 67(3):723–740.

Cunanan, K. M., Gonen, M., Shen, R., Hyman, D. M., Riely, G. J., Begg, C. B., and Iasonos, A. (2017a).
Basket trials in oncology: A trade-off between complexity and efficiency.
Journal of Clinical Oncology, 35(3):271–273.

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

Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2007).
Statistik.
Springer-Lehrbuch. Springer, Berlin Heidelberg New York, 5th edition.

Frederic, P. and Lad, F. (2008).
Two moments of the logitnormal distribution.
Communications in Statistics - Simulation and Computation, 37(7):1263–1269.

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.

Fujikawa, K., Teramukai, S., Yokota, I., and Daimon, T. (2020).
A Bayesian basket trial design that borrows information across strata based on the similarity between the posterior distributions of the response probability.
Biometrical Journal, 62(2):330–338.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004).
Bayesian Data Analysis.
Statistical Science Series. Chapman and Hall/CRC, 2nd edition.

Hartley, H. O. and Fitch, E. R. (1951).
A chart for the incomplete beta-function and the cumulative binomial distribution.
Biometrika, 38(3-4):423–426.

Held, L. and Bové, D. S. (2020).

Likelihood and Bayesian Inference.

Statistics for Biology and Health. Springer, Berlin, Heidelberg, 2nd edition.

Hobbs, B. P. and Landin, R. (2018).

Bayesian basket trial design with exchangeability monitoring.

Statistics in Medicine, 37(25):3557–3572.

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., Gervais, R., Elez-Fernandez, M. E., Italiano, A., Hofheinz, R. D., Hidalgo, M., Chan, E., Schuler, M., Lasserre, S. F., Makrutzki, M., Sirzen, F., Veronese, M. L., Tabernero, J., and Baselga, J. (2015).

Vemurafenib in multiple nonmelanoma cancers with braf v600 mutations.

New England Journal of Medicine, 373(8):726–36.

Jin, J., Liu, Q., Zheng, W., Shun, Z., Lin, T. T., Gao, L., and Dong, Y. (2020a).

A Bayesian method for the detection of proof of concept in early phase oncology studies with a basket design.

Statistics in Biosciences, 12(2):167–179.

Jin, J., Riviere, M.-K., Luo, X., and Dong, Y. (2020b).

Bayesian methods for the analysis of early-phase oncology basket trials with information borrowing across cancer types.

Statistics in Medicine, 39(25):3459–3475.

Johnson, N. L. (1949).

Systems of frequency curves generated by methods of translation.

Biometrika, 36(Pt. 1-2):149–76.

Joshi, Y. B. and Light, G. A. (2018).

Using EEG-guided basket and umbrella trials in psychiatry: A precision medicine approach for cognitive impairment in schizophrenia.

Frontiers in Psychiatry, 9(554).

Kaizer, A. M., Koopmeiners, J. S., and Hobbs, B. P. (2017).

Bayesian hierarchical modeling based on multisource exchangeability.
Biostatistics, 19(2):169–184.

Kalia, M. (2013).
Personalized oncology: recent advances and future challenges.
Metabolism, 62 Suppl 1:S11–4.

Kim, E. S., Herbst, R. S., Wistuba, I., Lee, J. J., Blumenschein, G. R., J., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., J., Gupta, S., Alden, C. M., Liu, S., Tang, X., Khuri, F. R., Tran, H. T., Johnson, B. E., Heymach, J. V., Mao, L., Fossella, F., Kies, M. S., Papadimitrakopoulou, V., Davis, S. E., Lippman, S. M., and Hong, W. K. (2011).
The battle trial: personalizing therapy for lung cancer.
Cancer Discovery, 1(1):44–53.

Kopp-Schneider, A., Wiesenfarth, M., Witt, R., Edelmann, D., Witt, O., and Abel, U. (2019).
Monitoring futility and efficacy in phase II trials with Bayesian posterior distributions-a calibration approach.
Biometrical Journal, 61(3):488–502.

Li, W., Chen, C., Li, X., and Beckman, R. A. (2017).
Estimation of treatment effect in two-stage confirmatory oncology trials of personalized medicines.
Statistics in Medicine, 36(12):1843–1861.

Li, W., Zhao, J., Li, X., Chen, C., and Beckman, R. A. (2019).
Multi-stage enrichment and basket trial designs with population selection.
Statistics in Medicine, 38(29):5470–5485.

Lieberman, G. J. and Owen, D. B. (1961).
Tables of the Hypergeometric Probability Distribution.
Stanford University Press, Stanford, CA.

Lin, R., Thall, P. F., and Yuan, Y. (2021).
A phase I-II basket trial design to optimize dose-schedule regimes based on delayed outcomes.

Bayesian Analysis, 16(1):179–202.

Liu, R., Liu, Z., Ghadessi, M., and Vonk, R. (2017).
Increasing the efficiency of oncology basket trials using a Bayesian approach.
Contemporary Clinical Trials, 63:67–72.

Lyu, J., Zhou, T., Yuan, S., Guo, W., and Ji, Y. (2020).
Muce: Bayesian hierarchical modeling for the design and analysis of phase 1b multiple expansion cohort trials.

Meyer, E. L., Mesenbrink, P., Dunger-Baldauf, C., Fülle, H.-J., Glimm, E., Li, Y., Posch, M., and König, F. (2020).
The evolution of master protocol clinical trial designs: A systematic literature review.
Clinical Therapeutics, 42(7):1330–1360.

Mitchell, T. J. and Beauchamp, J. J. (1988).
Bayesian variable selection in linear regression.
Journal of the American Statistical Association, 83(404):1023–1032.

Morita, S., Thall, P. F., and Müller, P. (2008).
Determining the effective sample size of a parametric prior.
Biometrics, 64(2):595–602.

Moscovich, A., Nadler, B., and Spiegelman, C. (2016).
On the exact Berk-Jones statistics and their p-value calculation.
Electronic Journal of Statistics, 10(2):2329–2354, 26.

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):123–34.

Novelli, M. (2021).
Information borrowing in phase II basket trials: a comparison of different designs.
Book of Short Papers SIS 2021, pages 572–577.

Park, J. J. H., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K., and Mills, E. J. (2019).
Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols.
Trials, 20(1):572.

Plummer, M. (2003).
Jags: A program for analysis of Bayesian graphical models using Gibbs sampling.
In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, pages 1–10.

Plummer, M. (2019).
rjags: Bayesian Graphical Models using MCMC.
R package version 4-10.

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.

Psioda, M. A., Xu, J., Jiang, Q., Ke, C., Yang, Z., and Ibrahim, J. G. (2019).
Bayesian adaptive basket trial design using model averaging.
Biostatistics, 22(1):19–34.

R Core Team (2021).
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria.

Redig, A. J. and Janne, P. A. (2015).
Basket trials and the evolution of clinical trial design in an era of genomic medicine.
Journal of Clinical Oncology, 33(9):975–977.

Robert, C. P., Chopin, N., and Rousseau, J. (2009).
Harold Jeffreys's theory of probability revisited.
Statistical Science, 24(2):141–172, 32.

RStudio Team (2021).

RStudio: Integrated Development Environment for R.

RStudio, PBC, Boston, MA.

Saville, B. R., Connor, J. T., Ayers, G. D., and Alvarez, J. (2014).

The utility of Bayesian predictive probabilities for interim monitoring of clinical trials.

Clinical Trials, 11(4):485–493.

Sebastiao, Y. V. and St. Peter, S. D. (2018).

An overview of commonly used statistical methods in clinical research.

Seminars in Pediatric Surgery, 27(6):367–374.

Simon, R. (1989).

Optimal two-stage designs for phase II clinical trials.

Controlled Clinical Trials, 10(1):1–10.

Simon, R., Geyer, S., Subramanian, J., and Roychowdhury, S. (2016).

The Bayesian basket design for genomic variant-driven phase II trials.

Seminars in Oncology, 43(1):13–18.

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):763–80.

Wickham, H. (2016).

ggplot2: Elegant Graphics for Data Analysis.

Springer-Verlag New York.

Woodcock, J. and LaVange, L. M. (2017).

Master protocols to study multiple therapies, multiple diseases, or both.

New England Journal of Medicine, 377(1):62–70.

Yang, R. and Berger, J. (1997).

A catalog of noninformative priors.

Duke University, Technical Report(97-42).

Zaslavsky, B. G. (2010).

Bayesian versus frequentist hypotheses testing in clinical trials with dichotomous and countable outcomes.

Journal of Biopharmaceutical Statistics, 20(5):985–997.

Zheng, H. and Wason, J. M. S. (2020).

Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy.

Biostatistics, doi: 10.1093/biostatistics/kxaa019.

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.

Zhou, T. and Ji, Y. (2020).

RoBoT: a robust Bayesian hypothesis testing method for basket trials.

Biostatistics, doi: 10.1093/biostatistics/kxaa005.

Publications

Methodological publications

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.

This publication covers the systematic approach towards basket trial designs via the categorization and the modular construction kit. It contains parts of Subsection 2.1.1, Section 3.1, Subsection 3.2.3, and Section 4.1. The manuscript has been written by myself but may contain comments and corrections from the co-authors.

Project-based publications

Liermann J., Naumann P., Hommertgen A., **Pohl M.**, Kieser M., Debus J., Herfarth K. (2020). Carbon ion radiotherapy as definitive treatment in non-metastasized pancreatic cancer: study protocol of the prospective phase II PACK-study. BMC Cancer, 20(1), 947. https://doi.org/10.1186/s12885-020-07434-8

Ades A. E., Brickley E. B., Alexander N., Brown D., Jaenisch T., Miranda-Filho, D. B., **Pohl M.**, Rosenberger K. D., Soriano-Arandes A., Thorne C., Ximenes R., de Araújo T., Avelino-Silva V. I., Bethencourt Castillo S. E., Borja Aburto V. H., Brasil P., Christie C., de Souza W. V., Gotuzzo H J. E., Hoen B., EC Zika Consortia Vertical Transmission Study Group (2020). Zika virus infection in pregnancy: a protocol for the joint analysis of the

prospective cohort studies of the ZIKAlliance, ZikaPLAN and ZIKAction consortia. BMJ Open, 10(12), e035307. https://doi.org/10.1136/bmjopen-2019-035307

Avelino-Silva V. I., Mayaud P., Tami A., Miranda M. C., Rosenberger K. D., Alexander N., Nacul L., Segurado A., **Pohl M.**, Bethencourt S., Villar L. A., Viana I., Rabello R., Soria C., Salgado S. P., Gotuzzo E., Guzmán M. G., Martínez P. A., López-Gatell H., Hegewisch-Taylor J., ZIKAlliance Clinical Study Group (2019). *Study protocol for the multicentre cohorts of Zika virus infection in pregnant women, infants, and acute clinical cases in Latin America and the Caribbean: the ZIKAlliance consortium.* BMC Infectious Diseases, 19(1), 1081. https://doi.org/10.1186/s12879-019-4685-9

Tisch M., Maier S., Preyer S., Kourtidis S., Lehnerdt G., Winterhoff S., Dalchow C. V., Mueller-Jenckel F., Sudhoff H. H., Schröder S., Koitschev A., Amrhein P., Bruchhage K. L., Leichtle A., Güldner C., Grulich-Henn J., Jensen K., **Pohl M.**, Plinkert P. K., Euteneuer S. (2020). *Balloon Eustachian Tuboplasty (BET) in Children: A Retrospective Multicenter Analysis.* Otology & neurotology : official publication of the American Otological Society, American Neurotology Society and European Academy of Otology and Neurotology, 41(7), e921–e933. https://doi.org/10.1097/MAO.0000000000002789

Kowalewski K. F., Schmidt M. W., Proctor T., **Pohl M.**, Wennberg E., Karadza E., Romero P., Kenngott H. G., Müller-Stich B. P., Nickel F. (2018). *Skills in minimally invasive and open surgery show limited transferability to robotic surgery: results from a prospective study.* Surgical endoscopy, 32(4), 1656–1667. https://doi.org/10.1007/s00464-018-6109-0

Mayer B., **Pohl M.**, Hummler H. D., Schmid M. B. (2017). *Cerebral oxygenation and desaturations in preterm infants - a longitudinal data analysis.* Journal of Neonatal-Perinatal Medicine, 10(3), 267–273. https://doi.org/10.3233/NPM-16124

Conference contributions

The presenting author is underlined.

Pohl M., Fischer D., Beyersmann J., Stucke-Straub K. An application of Bayesian regression models with covariates for dose finding in cancer phase I trials. *64. Biometrisches Kolloquium.* Presentation. March 2018, Frankfurt am Main, Germany

Pohl M., Krisam J., Kieser M., Congruencies and disparities between frequentist and Bayesian decision rules for futility and efficacy in basket trials. *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) e.V.*. Presentation. September 2019, Dortmund, Germany

Pohl M., Krisam J., Kieser M., Modular components in basket trials and connections among the applied tools. *42nd Annual Conference of the International Society for Clinical Biostatistics (ISCB 2021).* (virtual) Presentation. July 2021, Lyon, France

Pohl M., Krisam J., Kieser M., Key methodological innovations of selective Basket trial designs and their connections among each other. *32nd Conference of the Austro-Swiss Region (ROeS) of the International Biometric Society.* (virtual) Presentation. September 2021, Salzburg, Austria

Appendix A

Appendix - Additional results

A.1 Additional plots of correction factor κ

Plots of the correction factor κ are shown under different total number of patients and with different reference values p_0 in order to show the decline of κ with increasing p_0. The plots are equivalent to those shown in Section 3.2.1.

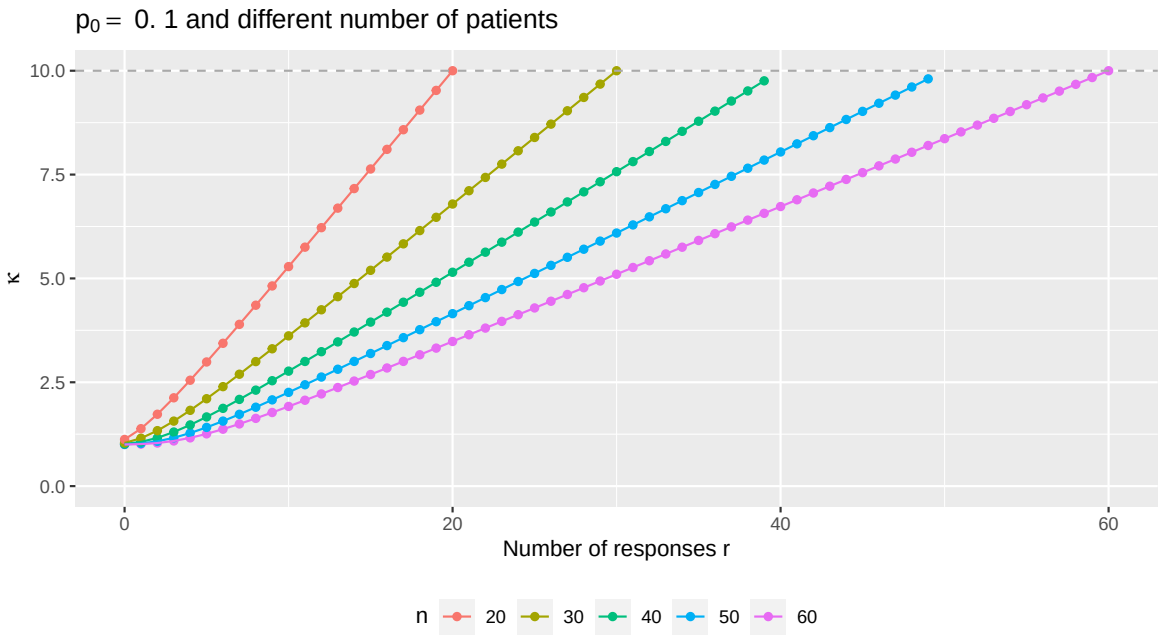


Figure A.1: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.10$.*

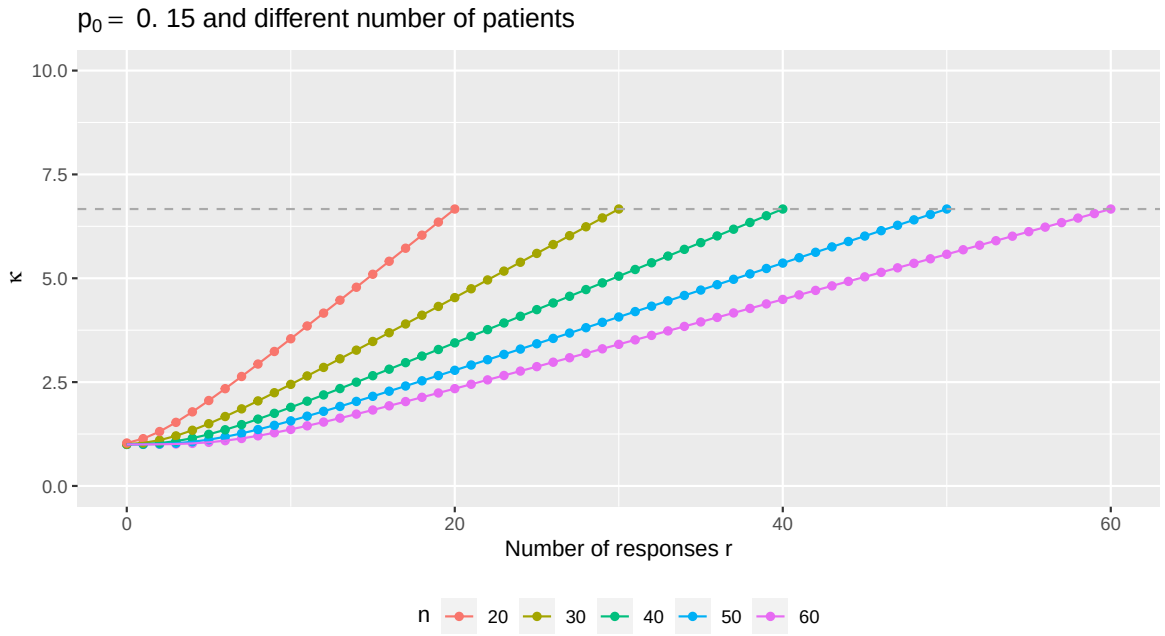


Figure A.2: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.15$.*

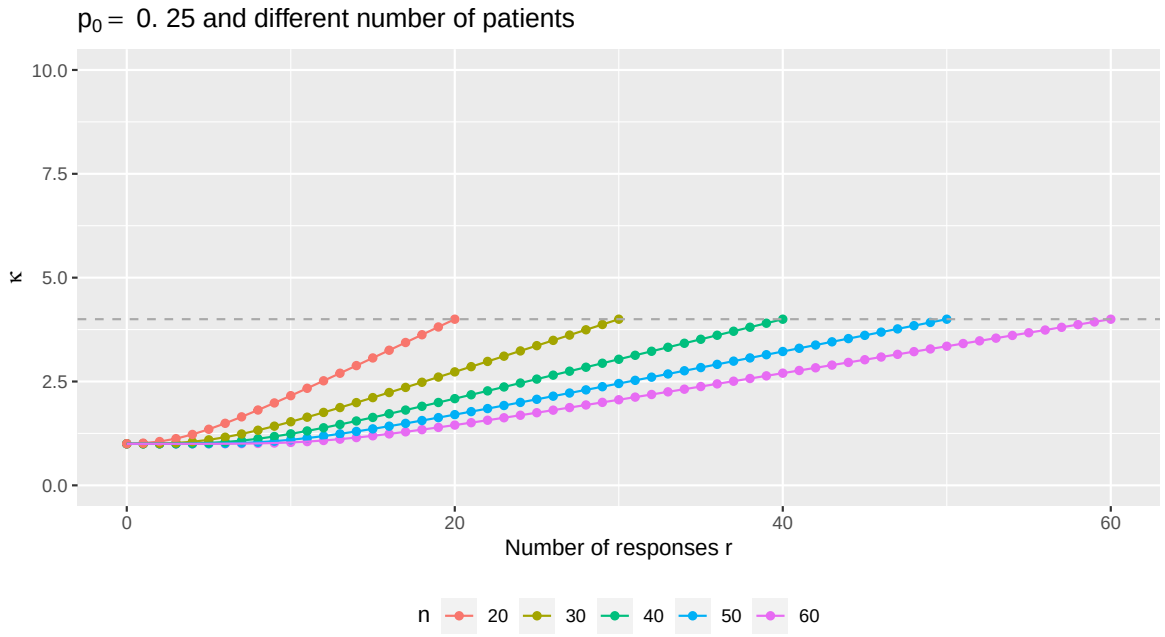


Figure A.3: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.25$.*

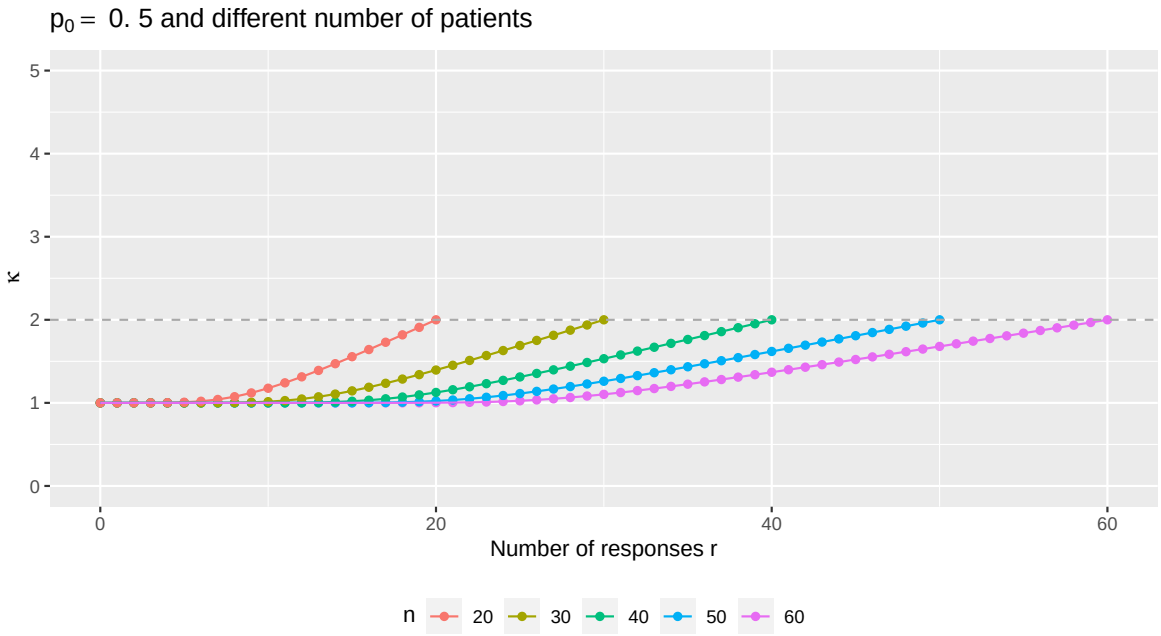


Figure A.4: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.50$.*

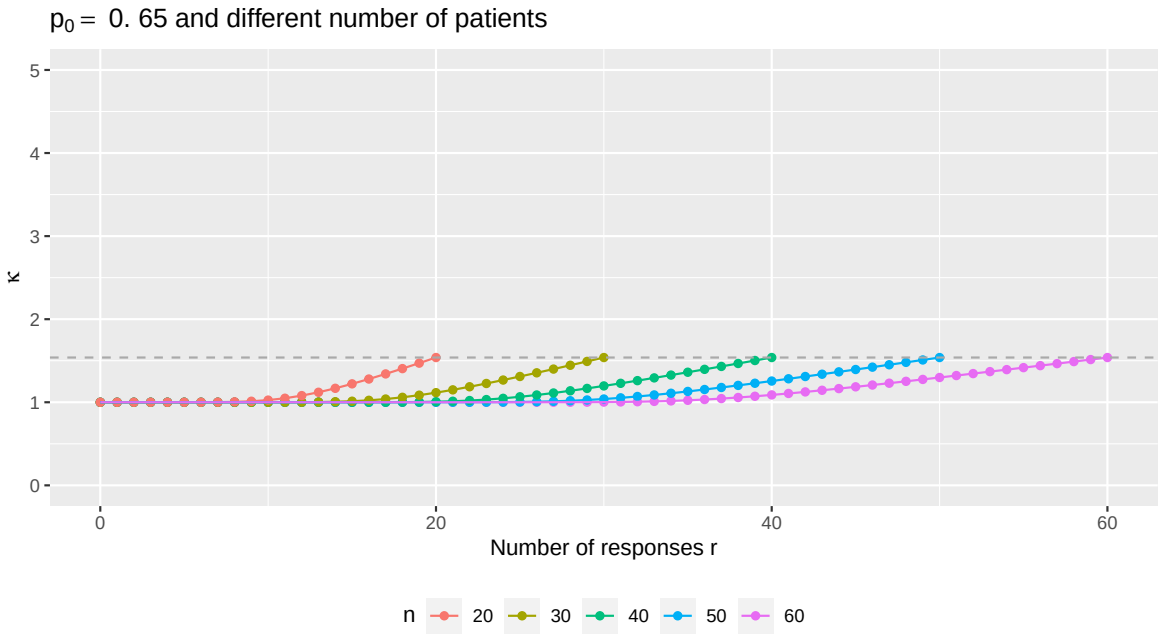


Figure A.5: *Correction factor κ for different total number of patients n and its monotonic increase in r up to the maximum κ_{max} which only depends on the reference value $p_0 = 0.65$.*

A.2 Additional plots of the absolute difference δ

Plots of the absolute difference δ in relation to the observed response rate $\frac{r}{n}$ are shown for different reference values p_0. The plots show that the difference is high in regions where discriminatory decisions with respect to the hypothesis are made. Moreover, they show that with increasing p_0 the absolute difference declines.

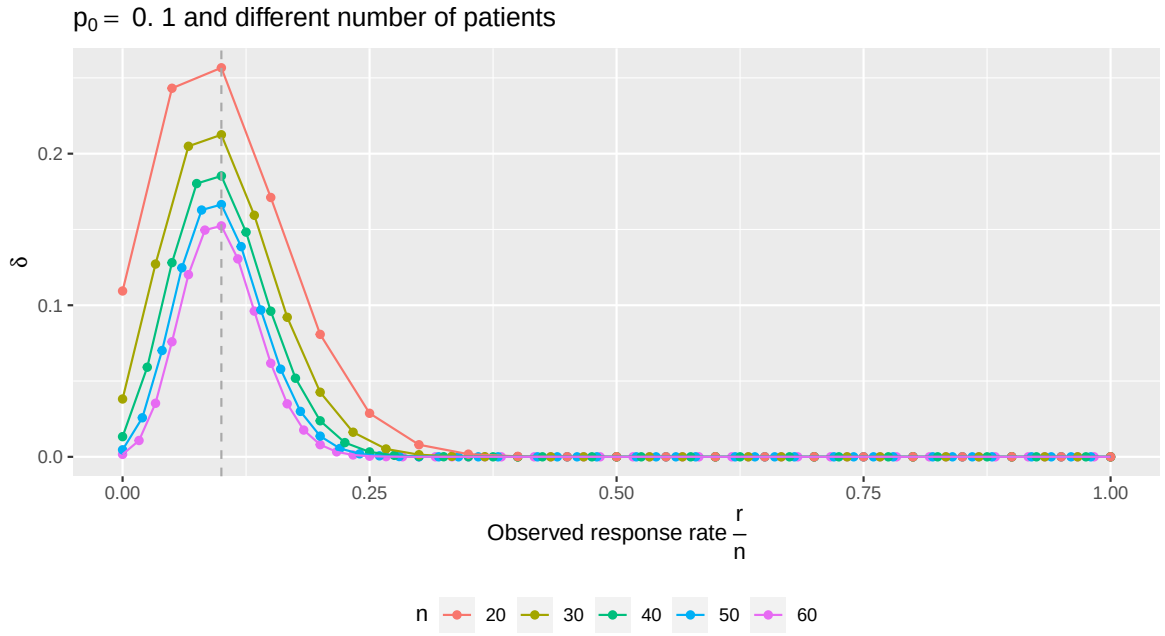


Figure A.6: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value* $p_0 = 0.10$.

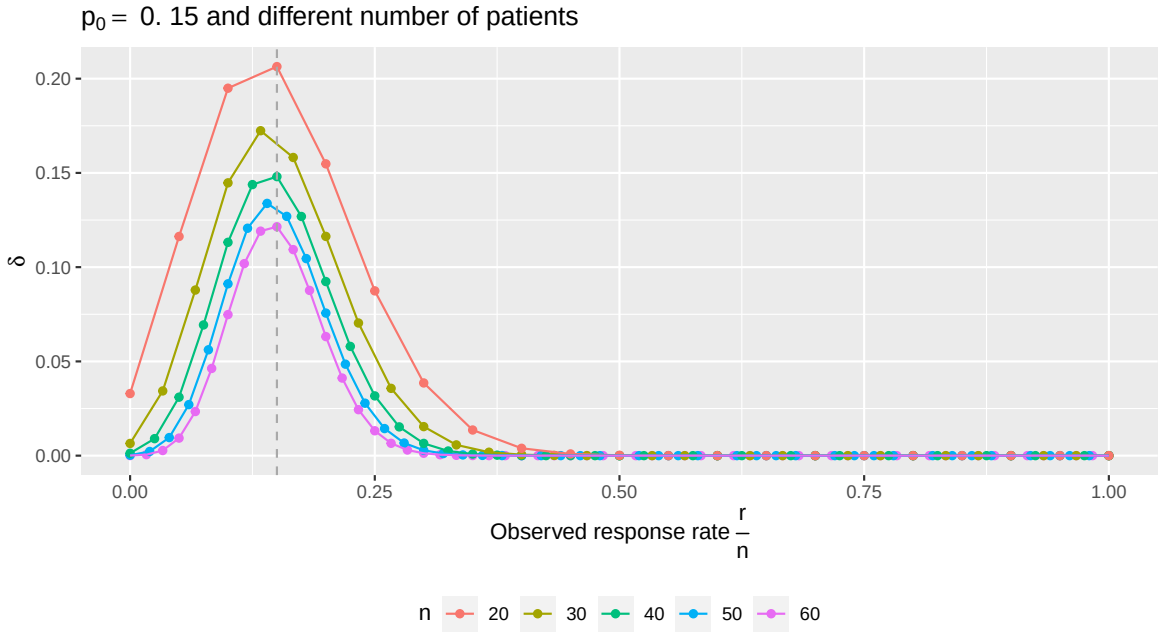


Figure A.7: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value* $p_0 = 0.15$.

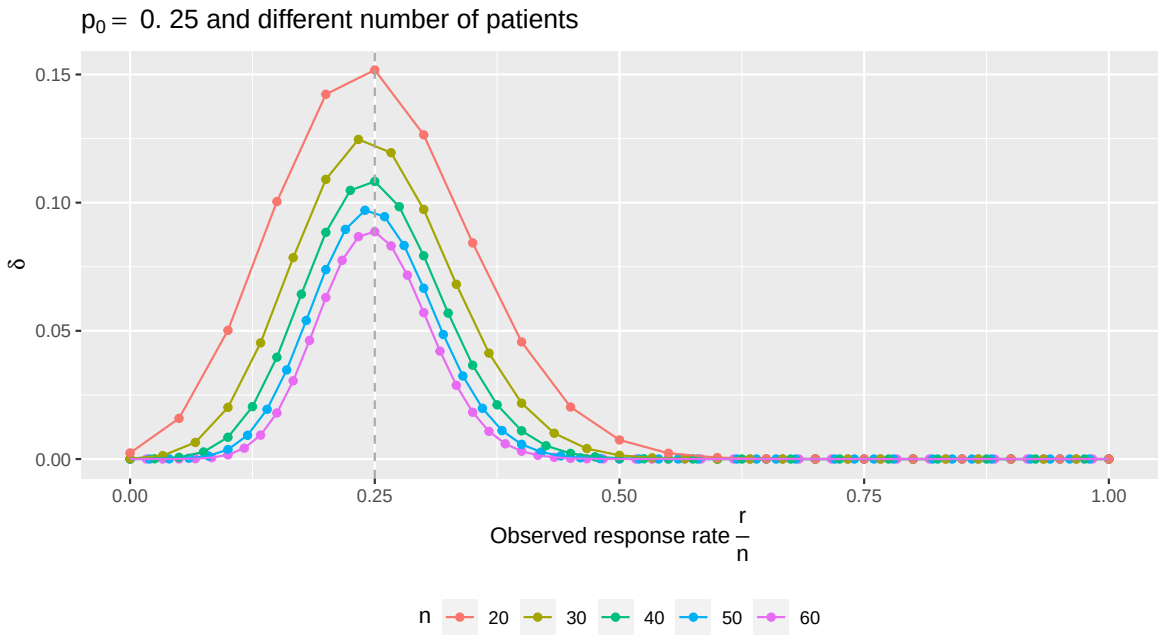


Figure A.8: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value* $p_0 = 0.25$.

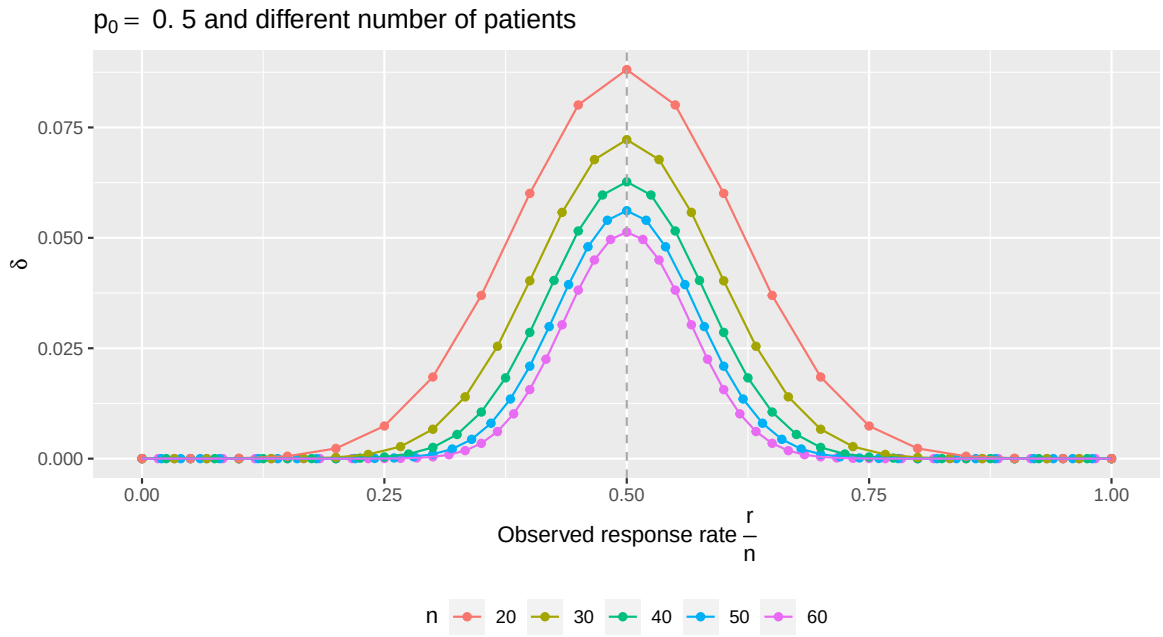


Figure A.9: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value $p_0 = 0.50$.*

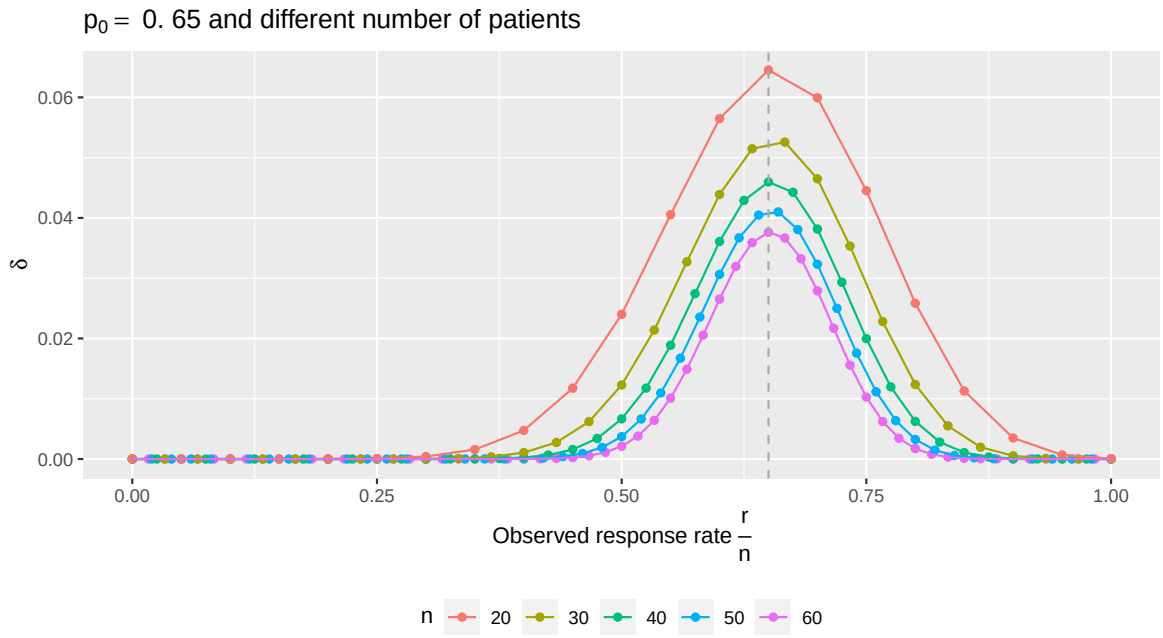


Figure A.10: *Absolute difference between $\mathcal{P}$ and $\mathcal{B}$ for different total number of patients n in relation to the observed response rate $\frac{r}{n}$. The dashed vertical line marks the reference value $p_0 = 0.65$.*

A.3 Additional posterior probabilities for the conversion of Cunanan et al. (2017b)

This section contains the tables with the posterior probabilities $\mathrm{P}[p_i > p_0|n, r]$ for r_{min} and $r_{min}-1$ under varying number of baskets in stage 2, denoted as i^*, in the heterogeneous path. The corresponding minimum number of responses r_{min} for every i^* are given in Section 3.2.2 in Table 3.3.

A.3.1 $i^* = 1$, $\dfrac{\alpha_s}{i^*} = \dfrac{0.07}{1} = 0.0700$, $r_{min} \in \{6, 7\}$

Table A.1: *Posterior probability $P[p_i > p_0|n, r]$ for the final analysis with the intention to declare a basket promising if at least r_{min} responses were observed. The number of patients per basket n among the $i^* = 1$ different baskets in stage 2 is varying.*

	n							
	18	19	20	21	22	23	24	25
r_{min}	0.9837	0.9781	0.9713	0.9886	0.9848	0.9801	0.9745	0.9679
$r_{min} - 1$	0.9463	0.9327	0.9173	0.9632	0.9537	0.9428	0.9305	0.9167

A.3.2 $i^* = 2$, $\dfrac{\alpha_s}{i^*} = \dfrac{0.07}{2} = 0.0350$, $r_{min} \in \{7, 8\}$

Table A.2: *Posterior probability $P[p_i > p_0|n, r]$ for the final analysis with the intention to declare a basket promising if at least r_{min} responses were observed. The number of patients per basket n among the $i^* = 2$ different baskets in stage 2 is varying.*

	n							
	18	19	20	21	22	23	24	25
r_{min}	0.9959	0.9941	0.9917	0.9886	0.9958	0.9941	0.9920	0.9894
$r_{min} - 1$	0.9837	0.9781	0.9713	0.9632	0.9848	0.9801	0.9745	0.9679

A.3.3 $i^* = 4$, $\frac{\alpha_s}{i^*} = \frac{0.07}{4} = 0.0175$, $r_{min} \in \{7, 8, 9\}$

Table A.3: *Posterior probability $P[p_i > p_0|n, r]$ for the final analysis with the intention to declare a basket promising if at least r_{min} responses were observed. The number of patients per basket n among the $i^* = 4$ different baskets in stage 2 is varying.*

	n							
	18	19	20	21	22	23	24	25
r_{min}	0.9959	0.9941	0.9980	0.9970	0.9958	0.9941	0.9979	0.9970
$r_{min} - 1$	0.9837	0.9781	0.9917	0.9886	0.9848	0.9801	0.9920	0.9894

A.3.4 $i^* = 5$, $\frac{\alpha_s}{i^*} = \frac{0.07}{5} = 0.0140$, $r_{min} \in \{7, 8, 9\}$

Table A.4: *Posterior probability $P[p_i > p_0|n, r]$ for the final analysis with the intention to declare a basket promising if at least r_{min} responses were observed. The number of patients per basket n among the $i^* = 5$ different baskets in stage 2 is varying.*

	n							
	18	19	20	21	22	23	24	25
r_{min}	0.9959	0.9987	0.9980	0.9970	0.9958	0.9985	0.9979	0.9970
$r_{min} - 1$	0.9837	0.9941	0.9917	0.9886	0.9848	0.9941	0.9920	0.9894

A.4 Requirement for logit-transformation of beta distribution

The transformation of a density function $f(x)$ by a differentiable function $g()$ requires that $g'(x) \neq 0$ for all $x \in [a, b]$ where $[a, b]$ is the space on which $f(x)$ is defined. In Subsection 3.3.1, the transformation function is $g(x) := logit(x)$ and the definition space of the density function of a beta distribution is $(0, 1)$.

The derivative $g'(x) = logit(x)' = log\left(\frac{x}{1-x}\right)$ is

$$\begin{aligned} g'(x) &= \frac{1-x}{x} \cdot \left(x(1-x)^{-1}\right)' = \frac{1-x}{x} \cdot \left((1-x)^{-1} + x(1-x)^{-2}(-1)\right) \\ &= \frac{1}{x} - \frac{1}{1-x} = \frac{1-x}{x(1-x)} - \frac{x}{x(1-x)} = \frac{1}{x-x^2} \end{aligned}$$

and for all $x \in (0, 1)$ it holds that $g'(x) \neq 0$.

Appendix B

Appendix - Implementations in R

B.1 Bayesian behaviour - JAGS code for posterior distributions

```
#-------------------------------------------------------------------------------
# quantiles of the posterior distribution
#-------------------------------------------------------------------------------
# data for posterior behavior
d <- data.frame(made = sort(rep(0:10,3)),
                attempts = c(rep(10,11*3)))
# - increased number of observations
# d <- data.frame(made = sort(rep(0:20,3)),
#                 attempts = c(rep(20,21*3)))

sigma2 <- sigma^2

# inits for MCMC sampling
inits <- list(list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 200618),
              list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 210618),
              list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 220618),
              list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 210621))

# data and prior distribution
dat = list(y = d$made,
           n = d$attempts,
           N = nrow(d),
           mu = mu,
```

```
24              sigma2 = sigma2)
25
26 #-----------------------------------------------
27 # logit-normal
28
29 lgn="
30 model {
31 prec = 1/sigma2
32
33 for(i in 1:N){
34     y[i] ~ dbinom(p[i], n[i])
35
36     p[i] = exp(theta[i])/(1+exp(theta[i]))
37     theta[i] ~ dnorm(mu, prec)
38 }
39 }
40 "
41
42 #-----------------------------------------------
43 # implement MCMC
44 model_lgn <- jags.model(textConnection(lgn),
45                         data = dat,
46                         n.chains = 4,
47                         inits = inits)
48
49 #-----------------------------------------------
50 # sample the indicator values(which.i) for theta
51 samples_lgn <- coda.samples(model = model_lgn,
52                             variable.names = c("p"),
53                             n.iter = 10^5)
54
55 quantiles_logitn <- summary(samples_lgn)$quantiles
56 rownames(quantiles_logitn) <- as.character(1:33)
57 xtable(quantiles_logitn , digits = 6)
58
59 #-----------------------------------------------
60 # beta
61
62 # robustness check leaving out 0 and 10
63 # d <- data.frame(made = sort(rep(1:9,3)),
64 #                 attempts = c(rep(10,9*3)))
65
66 dat = list(y = d$made,
67            n = d$attempts,
68            N = nrow(d),
```

```
            a = ml_a,
            b = ml_b)

# without adding T(0.000000000001,0.999999999999999) the model does not converge
bb="
model {
for(i in 1:N){
    p[i] ~ dbeta(a,b)T(0.000000000001,0.999999999999999)
    y[i] ~ dbinom(p[i], n[i])
}
}
"

#-----------------------------------------------
# implement MCMC
model_bb <- jags.model(textConnection(bb),
                       data = dat,
                       n.chains = 4,
                       inits = inits)

#-----------------------------------------------
# sample the indicator values(which.i) for theta
samples_bb <- coda.samples(model = model_bb,
                           variable.names = c("p"),
                           n.iter = 10^5)
quantiles_beta <- summary(samples_bb)$quantiles
rownames(quantiles_beta) <- as.character(1:33)
xtable(quantiles_beta, digits = 5)
```

B.2 Simulation study for comparison of BHM and HBB

```
################################################################################
#  Load packages and functions
################################################################################
library(rjags)
library(parallel)

logit <- function(p){
  log(p/(1-p))
}

expit <- function(x){
  exp(x)/(1+exp(x))
}

################################################################################
#  Definition of general variables
################################################################################
set.seed(100200)
#-------------------------------#
#    logit-normal
#-------------------------------#
# Prior
mu <- logit(.3)
sigma2 <- 7.8^2

# logit normal model
lgn="
model {

prec = 1/sigma2

for(i in 1:N){
    y[i] ~ dbinom(p[i], n[i])

    p[i] = exp(theta[i])/(1+exp(theta[i]))
    theta[i] ~ dnorm(mu, prec)
}
}
"

#-------------------------------#
#    BHM
#-------------------------------#
```

```
# Prior: sharing
mu_exp <- logit(0.3)
mu_sigma2 <- 7.8^2

bhm_shape <- 0.01
bhm_rate <- 0.10

# BHM model
bhm="
model {
  mu ~ dnorm(mu_exp, mu_prec)
  mu_prec = 1/mu_sigma2

  prec ~ dgamma(shape, rate)
  sigma2 = 1/prec

  for(i in 1:N){
      theta[i] ~ dnorm(mu, prec)

      p[i] = exp(theta[i])/(1+exp(theta[i]))
      y[i] ~ dbinom(p[i], n[i])
  }
}
"

#-----------------------------------#
#   HBB
#-----------------------------------#
# Prior: sharing
# position based on MLE estimate of expit transformed N(logit(0.1), 8^2)
# a <- 0.1377218
# b <- 0.1944923
# position based on MLE estimate of expit transformed N(logit(0.3), 7.8^2)
a <- 0.1558973
b <- 0.1786819

# chosen based on assumed data of r=3 (r=6 for n = 20) for all baskets
# and similar 2.5% and 97.5% quantiles of posterior and most important
# an at least wider range between 2.5% and 97.5% for HBB
hbb_shape <- 3
hbb_rate <- 0.12

# HBB model accounting for zeros
hbb="
model {
```

```
89
90    mu ~ dbeta(a,b)
91    eta ~ dgamma(shape, rate)
92
93    alpha = eta*mu
94    beta = eta*(1-mu)
95
96    for(i in 1:N){
97        p[i] ~ dbeta(alpha,beta)T(0.0000001,)
98        y[i] ~ dbinom(p[i], n[i])
99    }
100 }
101 "
102
103 #---------------------------------#
104 #   Functions
105 #---------------------------------#
106 data_sampler <- function(baskets_sample, p_sample, n_sample){
107   r_sample <- rbinom(n = baskets_sample,
108                      prob = p_sample,
109                      size = n_sample)
110   return(r_sample)
111 }
112
113 #---------------------------------#
114 #   Simulation characteristics
115 #---------------------------------#
116
117 # number of simulated studies
118 n_sim_studies <- 10000
119
120 # set seed for inital value of the chain - reproducibility
121 inits <- list(list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 200618),
122               list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 210618),
123               list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 220618),
124               list(".RNG.name" = "base::Wichmann-Hill", ".RNG.seed" = 210621))
125
126 # number of basktes in simulation
127 n_baskets_sample <- 6
128
129 # number of patients/observations per basket
130 n_sample <- c(rep(20, n_baskets_sample))
131
132 # assumed response rates to sample data
133 p_sample_list <- list(rep(0.3, n_baskets_sample),
```

```
                        rep(0.2, n_baskets_sample),
                        rep(0.1, n_baskets_sample),

                        c(0.1, rep(0.3, n_baskets_sample - 1)),
                        c(0.1, 0.1, rep(0.3, n_baskets_sample - 2)),
                        c(0.1, 0.1, 0.1, rep(0.3, n_baskets_sample - 3)),

                        c(0.1, 0.20, rep(0.3, n_baskets_sample - 2)),
                        c(0.1, 0.15, rep(0.3, n_baskets_sample - 2)),
                        c(0.1, 0.15, 0.2, rep(0.3, n_baskets_sample - 3)),

                        c(seq(0.05, 0.3, by = 0.05)),
                        c(seq(0.18, 0.28, by = 0.02)),
                        c(seq(0.18, 0.23, by = 0.01))
)

################################################################################
#  Simulate posterior distributions with models
################################################################################

#-------------------------------------------------------------------------------
# Sample data for each basket under given p and n
sample_data <- function(p_sample){
  p_sample <- p_sample

  # set counter
  counter_study <- 0

  while(counter_study < n_sim_studies){

    # increase counter
    counter_study <- counter_study + 1

    r_sample <- data_sampler(baskets_sample = n_baskets_sample,
                             p_sample = p_sample,
                             n_sample = n_sample)

    # storage for simulated number of responses
    if (counter_study == 1){
      storage_r <- r_sample
    }else{
      storage_r <- rbind(storage_r, r_sample)
    }
  }
  colnames(storage_r) <- p_sample
```

```
179   rownames(storage_r) <- 1:n_sim_studies
180
181   # reset counter to 0
182   if (counter_study == n_sim_studies){
183     counter_study <- 0
184   }
185
186   return(storage_r)
187 }
188
189 #-------------------------------------------------------------------------------
190 # Sample posterior distributions for individual evaluation, BHM and HBB
191
192 #-------------------------------------
193 # function for analytic beta-binomial
194
195 # all possible responses
196 r_analytical <- c(0:n_sample[1])
197
198 # analytic mean posterior probability to exceed the sample probability
199 # in individual beta-binomial model
200 ana_bb <- function(p_sample_scenario, p0=FALSE){
201   sapply(p_sample_scenario, function(x){
202     lhood <- dbinom(r_analytical, n_sample[1], x)
203     if(p0==F){
204       post <- 1 - pbeta(x, a + r_analytical, b + n_sample[1] - r_analytical)
205     }else{
206       post <- 1 - pbeta(p0, a + r_analytical, b + n_sample[1] - r_analytical)
207     }
208     return(sum(lhood*post))
209   })
210 }
211
212 #-------------------------------------
213 # function for logitnormal model
214 sim_lgn <- function(storage_r){
215   # reset counter
216   counter_study <- 0
217
218   # add the already sampled data
219   storage_r <- storage_r
220   p_sample <- as.numeric(colnames(storage_r))
221
222   while(counter_study < n_sim_studies) {
223
```

```
224     # increase counter
225     counter_study <- counter_study + 1
226
227     #-----------------------------------------
228     # JAGS to sample posterior probabilities
229
230     #.....................
231     #   logit-normal
232     #.....................
233     dat_lgn = list(y = storage_r[counter_study,],
234                    n = n_sample,
235                    N = n_baskets_sample,
236                    mu = mu,
237                    sigma2 = sigma2)
238
239     model_lgn <- jags.model(textConnection(lgn),
240                             data = dat_lgn,
241                             n.chains = 4,
242                             inits = inits)
243
244     samples_lgn <- coda.samples(model = model_lgn,
245                                 variable.names = c("p"),
246                                 n.iter = 10000)
247
248     #-----------------------------------------
249     # Analysis of the posterior samples with respect to assumed p
250
251     # combine the chains from JAGS sampling
252     combined_chains <- do.call(rbind, samples_lgn)
253
254     # calculate posterior probability to exceed sampling probability
255     sim_post_prob <- sapply(1:n_baskets_sample,
256                             function(x){
257                               sum(combined_chains[,x] > p_sample[x])/dim(combined_
                                      chains)[1]
258                             })
259
260     # calculate posterior probability to exceed null value 0.1
261     sim_post_prob_null <- sapply(1:n_baskets_sample,
262                                  function(x){
263                                    sum(combined_chains[,x] > 0.1)/dim(combined_chains)
                                         [1]
264                                  })
265
266     # storage for simulated posterior probabilities
```

```
267     if (counter_study == 1){
268       storage_post_prob <- sim_post_prob
269       storage_post_prob_null <- sim_post_prob_null
270     }else{
271       storage_post_prob <- rbind(storage_post_prob, sim_post_prob)
272       storage_post_prob_null <- rbind(storage_post_prob_null, sim_post_prob_null)
273     }
274   }
275
276   # names for col and row of result table
277   colnames(storage_post_prob) <- p_sample
278   rownames(storage_post_prob) <- 1:n_sim_studies
279   colnames(storage_post_prob_null) <- p_sample
280   rownames(storage_post_prob_null) <- 1:n_sim_studies
281   return(list(storage_post_prob,
282               storage_post_prob_null))
283 }
284
285 #-----------------------------------
286 # function for BHM
287 sim_bhm <- function(storage_r){
288
289   # reset counter
290   counter_study <- 0
291
292   # add the already sampled data
293   storage_r <- storage_r
294   p_sample <- as.numeric(colnames(storage_r))
295
296   while(counter_study < n_sim_studies) {
297
298     # increase counter
299     counter_study <- counter_study + 1
300
301     #-----------------------------------------
302     # JAGS to sample posterior probabilities
303
304     #.....................
305     #       BHM
306     #.....................
307     dat_bhm = list(y = storage_r[counter_study,],
308                    n = n_sample,
309                    N = n_baskets_sample,
310                    mu_exp = mu_exp,
311                    mu_sigma2 = mu_sigma2,
```

```
312                     shape = bhm_shape,
313                     rate = bhm_rate)
314
315     model_bhm <- jags.model(textConnection(bhm),
316                             data = dat_bhm,
317                             n.chains = 4,
318                             inits = inits)
319
320     samples_bhm <- coda.samples(model = model_bhm,
321                                 variable.names = c("p"),
322                                 n.iter = 10000)
323
324     #-----------------------------------------
325     # Analysis of the posterior samples with respect to assumed p
326
327     # combine the chains from JAGS sampling
328     combined_chains <- do.call(rbind, samples_bhm)
329
330     # calculate posterior probability to exceed sampling probability
331     sim_post_prob <- sapply(1:n_baskets_sample,
332                             function(x){
333                               sum(combined_chains[,x] > p_sample[x])/dim(combined_
                                      chains)[1]
334                             })
335
336     # calculate posterior probability to exceed null value 0.1
337     sim_post_prob_null <- sapply(1:n_baskets_sample,
338                                  function(x){
339                                    sum(combined_chains[,x] > 0.1)/dim(combined_chains)
                                         [1]
340                                  })
341
342     # storage for simulated posterior probabilities
343     if (counter_study == 1){
344       storage_post_prob <- sim_post_prob
345       storage_post_prob_null <- sim_post_prob_null
346     }else{
347       storage_post_prob <- rbind(storage_post_prob, sim_post_prob)
348       storage_post_prob_null <- rbind(storage_post_prob_null, sim_post_prob_null)
349     }
350   }
351
352   # names for col and row of result table
353   colnames(storage_post_prob) <- p_sample
354   rownames(storage_post_prob) <- 1:n_sim_studies
```

```
  colnames(storage_post_prob_null) <- p_sample
  rownames(storage_post_prob_null) <- 1:n_sim_studies
  return(list(storage_post_prob,
              storage_post_prob_null))
}

#-------------------------------------
# function for HBB
sim_hbb <- function(storage_r){

  # reset counter
  counter_study <- 0

  # add the already sampled data
  storage_r <- storage_r
  p_sample <- as.numeric(colnames(storage_r))

  while(counter_study < n_sim_studies) {

    # increase counter
    counter_study <- counter_study + 1

    #-----------------------------------------
    # JAGS to sample posterior probabilities

    #.....................
    #       HBB
    #.....................
    dat_hbb = list(y = storage_r[counter_study,],
                   n = n_sample,
                   N = n_baskets_sample,
                   a = a,
                   b = b,
                   shape = hbb_shape,
                   rate = hbb_rate)

    model_hbb <- jags.model(textConnection(hbb),
                            data = dat_hbb,
                            n.chains = 4,
                            inits = inits)

    samples_hbb <- coda.samples(model = model_hbb,
                                variable.names = c("p"),
                                n.iter = 10000)

```

```
400     #-----------------------------------------------
401     # Analysis of the posterior samples with respect to assumed p
402
403     # combine the chains from JAGS sampling
404     combined_chains <- do.call(rbind, samples_hbb)
405
406     # calculate posterior probability to exceed sampling probability
407     sim_post_prob <- sapply(1:n_baskets_sample,
408                             function(x){
409                               sum(combined_chains[,x] > p_sample[x])/dim(combined_
                                    chains)[1]
410                             })
411
412     # calculate posterior probability to exceed null value 0.1
413     sim_post_prob_null <- sapply(1:n_baskets_sample,
414                                  function(x){
415                                    sum(combined_chains[,x] > 0.1)/dim(combined_chains)
                                         [1]
416                                  })
417
418     # storage for simulated posterior probabilities
419     if (counter_study == 1){
420       storage_post_prob <- sim_post_prob
421       storage_post_prob_null <- sim_post_prob_null
422     }else{
423       storage_post_prob <- rbind(storage_post_prob, sim_post_prob)
424       storage_post_prob_null <- rbind(storage_post_prob_null, sim_post_prob_null)
425     }
426   }
427
428   # names for col and row of result table
429   colnames(storage_post_prob) <- p_sample
430   rownames(storage_post_prob) <- 1:n_sim_studies
431   colnames(storage_post_prob_null) <- p_sample
432   rownames(storage_post_prob_null) <- 1:n_sim_studies
433   return(list(storage_post_prob,
434               storage_post_prob_null))
435 }
436
437 #-------------------------------------------------------------------------------
438 # Simulate different data scenarios
439 system.time(
440   raw_results <- mclapply(p_sample_list, function(x){
441
442     # simulate each data scenario
```

```
443     set.seed(124578)
444     storage_r <- sample_data(p_sample = x)
445
446     results_lgn <- sim_lgn(storage_r)
447     results_bhm <- sim_bhm(storage_r)
448     results_hbb <- sim_hbb(storage_r)
449
450     # return the results
451     ret_list <- list(storage_r = storage_r,
452                      results_lgn = results_lgn[[1]],
453                      results_bhm = results_bhm[[1]],
454                      results_hbb = results_hbb[[1]],
455
456                      results_lgn_null = results_lgn[[2]],
457                      results_bhm_null = results_bhm[[2]],
458                      results_hbb_null = results_hbb[[2]])
459     return(ret_list)
460   }, mc.cores = 12)
461 )
462
463 ##############################################################################
464 #  Analyze and store the results
465 ##############################################################################
466 # underlying simulation parameters
467 assumed_parameters <- list(n_sim_studies = n_sim_studies,
468                            n_baskets_sample = n_baskets_sample,
469                            n_sample = n_sample,
470                            p_sample_list = p_sample_list,
471                            mu = mu,
472                            sigma2 = sigma2,
473                            mu_exp = mu_exp,
474                            mu_sigma2 = mu_sigma2,
475                            bhm_shape = bhm_shape,
476                            bhm_rate = bhm_rate,
477                            a = a,
478                            b = b,
479                            hbb_shape = hbb_shape,
480                            hbb_rate = hbb_rate
481 )
482
483
484 # - Mean of P[p_i > p]
485 results_means <- lapply(raw_results, function(x){
486   mean_prob <- rbind(ana_bb(as.numeric(colnames(x$results_bhm))),
487                      colMeans(x$results_lgn),
```

```
488                       colMeans(x$results_bhm),
489                       colMeans(x$results_hbb))
490   rownames(mean_prob) <- c("analytic BB", "lgn", "BHM", "HBB")
491   return(mean_prob)
492 })
493 results_means
494
495 # - Mean of P[p_i > p0] - againts NULL value
496 results_means_null <- lapply(raw_results, function(x){
497   mean_prob_null <- rbind(ana_bb(as.numeric(colnames(x$results_bhm)), p0 = 0.1),
498                           colMeans(x$results_lgn_null),
499                           colMeans(x$results_bhm_null),
500                           colMeans(x$results_hbb_null))
501   rownames(mean_prob_null) <- c("analytic BB", "lgn_null", "BHM_null", "HBB_null")
502   return(mean_prob_null)
503 })
504 results_means_null
505
506 # - Mean responses of simulated data
507 results_sim_data <- lapply(raw_results, function(x){
508   return(colMeans(x$storage_r)
509   )
510 })
511 results_sim_data
512
513 #-------------------------------------#
514 # Save the data externally
515
516 # n = 10, logit(0.1)
517 # save(assumed_parameters,
518 #      results_means,
519 #      results_means_null,
520 #      results_sim_data,
521 #      file = "diss/sim_results/n10_logit01.RData")
```

Curriculum Vitae

Personal information

Moritz Pohl, born on the 2nd of November 1991 in Esslingen am Neckar, Germany

Education

Heidelberg University - Doctoral student (Dr. sc. hum.)	since 07/2017
Ulm University - Master of Science: Mathematical Biometry	10/2014 - 05/2017
Ulm University - Bachelor of Science: Mathematical Biometry	10/2011 - 08/2014
Schlossgymnasium Kirchheim unter Teck - A-Level (Abitur)	09/2002 - 06/2011

Professional experience

Heidelberg University - University Hospital since 07/2017
Research Fellow at Institute of Medical Biometry

Boehringer Ingelheim Pharma GmbH & Co. KG 04/2016 - 04/2017
Intern and working student at Clinical Statistics

Deutsche Boerse Group - Eurex Clearing 07/2015 - 10/2015
Intern at Clearing Supervision

AbbVie Deutschland GmbH & Co. KG 08/2013 - 10/2013
Intern at Data and Statistical Sciences

Acknowledgments

I like to thank my supervisor Prof. Dr. Meinhard Kieser for the possibility to write this thesis and for his scientific and general support whenever needed. Moreover, I am thankful for my colleagues at the Institute of Medical Biometry. The harmonic atmosphere and the cooperative spirit in our offices was truly helpful while working on this thesis.

Special thanks and gratitude goes to my girlfriend for her constant and unconditional support and for her encouragement in all situations. Finally, I like to thank my family for accompanying me throughout all my life, without their support I would not have come to this point. I am deeply grateful and thank you all very much.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema *Systematic approach and investigation of the statistical tools in basket trial designs* 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