Categories, components, and techniques in a modular construction of basket trials for application and further research

Moritz Pohl | Johannes Krisam | Meinhard Kieser

Institute of Medical Biometry and Informatics, Medical Biometry, University of Heidelberg, Heidelberg, Germany

Correspondence
Moritz Pohl, Institute of Medical Biometry and Informatics, Medical Biometry, University of Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany.
Email: pohl@imbi.uni-heidelberg.de

Abstract
Basket trials have become a virulent topic in medical and statistical research during the last decade. The core idea of them is to treat patients, who express the same genetic predisposition—either personally or their disease—with the same treatment irrespective of the location of the disease. The location of the disease defines each basket and the pathway of the treatment uses the common genetic predisposition among the baskets. This opens the opportunity to share information among baskets, which can consequently increase the information of the basket-wise response with respect to the investigated treatment. This further allows dynamic decisions regarding futility and efficacy of individual baskets during the ongoing trial. Several statistical designs have been proposed on how a basket trial can be conducted and this has left an unclear situation with many options. The different designs propose different mathematical and statistical techniques, different decision rules, and also different trial purposes. This paper presents a broad overview of existing designs, categorizes them, and elaborates their similarities and differences. A uniform and consistent notation facilitates the first contact, introduction, and understanding of the statistical methodologies and techniques used in basket trials. Finally, this paper presents a modular approach for the construction of basket trials in applied medical science and forms a base for further research of basket trial designs and their techniques.

Key words
basket trial, Bayes, frequentist, overview, precision medicine

1 | INTRODUCTION

In the last decade, the focus on personalized medicine has been a major interest of research and has currently reached just a starting point for future developments (e.g., Ashley, 2016; Jackson & Chester, 2015). Especially indications in oncology pose a variety of treatment options as the knowledge about cancer diseases and the unique characteristics of each patient become more prevalent and relevant for the choice of treatment. Based on the patient’s history, his or
her biomarkers, including the genetic information of the patient and/or the cancer, a tailored treatment might lessen the patient’s burden and improve his lifetime. The individual treatment choice is in general supposed to be made on objective patient characteristics. The need to tackle cancer is obvious as one out of six women and one out of five men are expected to develop cancer once throughout their lifetime and the worldwide number of 9.6 million deaths and 18.1 million new diagnoses of cancer in 2018 underline the threat of life to all people (Bray et al., 2018). Thus, the advances made in medical research demand also progress in statistical methods and designs of clinical trials to deliver efficacious personalized treatment to patients in medical need.

This progress is displayed by master protocols which have evolved as a generic term that includes basket, umbrella and platform trials. Although the nomenclature of each of the three trial designs is not uniquely defined, Woodcock and LaVange (2017) proposed to consider basket trials as studies with a single therapy for multiple diseases or subtypes and umbrella trials as studies with multiple therapies for one single disease. Platform trials are dynamic extensions of basket or umbrella trials and serve as what their name suggests: a platform to add and withdraw treatments (for umbrella trials) and diseases (for basket trials) during an ongoing trial (Collignon et al., 2020). The idea behind each of the master protocols is to evaluate the effect of personalized treatment, where the treatment is linked with the characteristics (e.g., genetics) of the disease and the patient respectively. Master protocols have become a virulent field of research in recent years as Park et al. (2019) and Meyer et al. (2020) showed in their systematic literature reviews. They both reported an exponential growth of publications which contain planned and conducted trials, review articles and methodological research, moreover they expect that the peak has not been reached yet. The majority of publications concentrate on scenarios from oncology; however, also research regarding ophthalmology, Alzheimer’s disease, or Ebola is conducted with master protocols. Also both indicate that basket trials are more dominant in practical application and methodological research than umbrella and platform trials.

The evolution of basket trials has brought with itself a number of trial designs and statistical methodologies. The designs cover trials from early phases of clinical development up to confirmatory studies. The methodologies present different techniques to gain advantage from the known, or at least assumed, linkage between the patients’ predisposition and the administered treatment. They refer to different scales of the primary endpoint but mostly focus on a binary response information because this outcome is the most practical one for early response detection in exploratory settings. The large amount of designs and methodologies has fabricated an unclear situation with many options in a very dynamic field of research. The goal of this work is to give an orderly approach to existing techniques, to present the existing methodologies in a consistent notation and to underline the modularity of basket trials. Those aspects will facilitate to draw connections between techniques, they will serve as a comprehensive overview for researchers getting in touch with available methods and will create a general base for further research on basket trials. This research is needed to keep up with advances in medical research and to deliver the required progress in trial designs and the underlying biometrical methods to prove the effectiveness of innovative and needed treatments.

In literature, the diseases or subtypes in a basket trial are denoted by various names, for example subpopulations, indications or strata, and they are also called baskets (e.g., in Chu & Yuan, 2018a; Cunanan et al., 2017b; Psioda et al., 2019). Throughout this paper, we apply the latter wording, such that a basket trial consists of several baskets where each basket contains patients with the same predisposition and the same disease. The variation in terminology must be kept in mind when consulting the original literature.

This work is structured as follows. In Section 2, we present the general components of a basket trial design for a modular construction. Also we elaborate on the workflow and different categories of basket trials. Section 3 is the main part of this work and we present the methodologies and techniques of all components of a basket trial in a consistent notation and draw connections between the methods. In Section 4, we present a practical proposal how to facilitate the planning and implementation of a basket trial in applied medical research from a statistician’s perspective. In Section 5, specific characteristics of basket trials, their components and methodologies are finally discussed.

2 | MODULARITY, WORKFLOW, AND CATEGORIES OF BASKET TRIALS

Basket trials mainly consist of four elementary design components. The first component is the sharing of information among baskets. The information sharing is the core element of basket trials because it utilizes the common genetic predisposition in each basket to increase the power to detect effective baskets. It is applied either on a regular basis before decisions are made or only occasionally at predefined nodes of the trial. Information sharing is mandatory at least once, because it reflects the general idea of a basket trial. The futility assessment is the second component. Its purpose is to save scarce resources in form of budget, workforce, and most importantly prevent patients from treatment with futile
treatments. Interim futility assessments are optional and can either be applied once or several times throughout the trial. The third component is also optional and is the interim efficacy assessment. It enables to stop recruitment to baskets with convincing evidence in favor of the novel treatment. The intention behind an efficacy stop is to quickly proceed with a confirmatory trial in case the basket trial is of exploratory nature or, in case of a confirmatory basket trial, to request accelerated market approval leading to a broad accessibility of the treatment to patients in need. An efficacy assessment can be conducted once or several times. It can take place at the same time as the futility assessment but does not necessarily have to. The fourth design component of a basket trial is the mandatory final analysis of the baskets. Depending on the goal of the trial, decisions regarding the efficacy of the treatment among the baskets are made. The first and the fourth component define a basket trial. Without doubt the sharing of information among the baskets is the key element as it reflects the medical assumption of similar treatment effects among the baskets justified by the common predisposition. The interim assessments of futility and efficacy are optional features of the trial design with advantages and disadvantages regarding sample size, ethics, and type 1 and 2 error rate control.

So far proposed trial designs use different statistical tools for each of the four components. Therefore, a study team has the option to choose trial components and the respective tools from a broad stock of methods to implement a new basket trial. The tools for each component can come from different publications and designs. Hence, a basket trial has a modular structure with two obligatory components (sharing of information, final analysis) and two optional components (interim assessment of futility and of efficacy). Therefore modular basket trials which apply sharing with a tool from one design, interim (futility and/or efficacy) analysis from another, and final analysis with a tool from a third design are practically feasible. The modular approach allows to create a trial well suited for the underlying medical and practical conditions with already existing tools. Each tool has its own technical properties. It is therefore important to investigate and understand the tools applied in every of the four elementary components of a basket design. The technical aspects and published tools for each component will be further investigated in Section 3.

The workflow of a basket trial depends on the numerous combinations for the arrangement of the four components. Apart from the obligatory final analysis and at least one node for the sharing of information, no further restrictions are given. The most frequent arrangement schedules information sharing before assessing futility and efficacy at interim analysis. This allows decisions being made on the basis of all observed data and to best of one’s knowledge. The vast majority of known basket designs employs this strategy. However, for example Liu et al. (2017) applies a futility assessment before the data are shared. This, at first sight counterintuitive step underlines the flexibility of how the components can be aligned. The scheme of possible workflows is presented in Figure A1 in the Appendix.

Basket trials can be categorized mainly by their purpose and their mathematical tools. The purpose of most basket trials is early detection of evidence for possible efficacy. Therefore they are mainly conducted as phase II trials to gather information to initiate large phase III trials in promising baskets. Jin et al. (2020a) explicitly called their design a proof-of-concept (PoC) approach for early development with the goal to state that the treatment works in at least one indication. Other authors (Asano & Hirakawa, 2020; Chen & Lee, 2019, 2020; Berry et al., 2013; Chu & Yuan, 2018b; Psioda et al., 2019; Chu & Yuan, 2018a; Neuenschwander et al., 2016; Fujikawa et al., 2020; Jin et al., 2020b; Hobbs & Landin, 2018; Lyu et al., 2020; Zhou & Ji, 2020; Simon et al., 2016; Zheng & Wason, 2019) refer to early clinical development in phase II with the intention to detect indications that work or at least present promising efficacy results. On the other hand (Chen et al., 2016; Li et al., 2017) propose designs for confirmatory phase III basket trials aiming to achieve approval for the treatment in several indications. The design of Chen et al. (2016) strictly concentrates on controlling the type I error (T1E) and does so by using frequentist statistics. It is accompanied by Beckman et al. (2016) who present the general challenges of the design, the conduct, and the interpretation of a confirmatory basket trial, referring to the design of Chen et al. (2016). Most of the proposed designs for early phases apply Bayesian techniques to analyze and conduct the trial. Still there are frequentist designs for the early phase (Cunanan et al., 2017b; Zhou et al., 2019) and there are also designs which use a combination of both techniques (Liu et al., 2017). Hence, basket designs can be categorized by their purpose (early phase exploratory, late phase confirmatory) and by the applied statistical techniques (frequentist, Bayesian, combination of both) to conduct and analyze the trials.

3  |  ANALYSIS OF BASKET TRIALS

The statistical analysis of a basket trial is defined by the combination of the individual analyses in each component. We present the technical aspects of each analysis tool, how they are further evolved to new techniques, and selective connections between each other in the following subsections. For ease of understanding, a consistent and common notation is initiated and applied.
The index \( i \) assigns a variable to basket \( i \). If no index is used, the variable represents the global parameter value that is valid for all baskets. If additional indices are used, they are explicitly defined in advance. Variable \( n \) describes the number of patients, \( r \) denotes the number of responses for the binary outcome and \( p \) describes the response rate. The set of all so far observed data is represented by \( D \) and the information from basket \( i \) is denoted by \( D_i := \{ n_i, r_i \} \). For transformations of \( p \) we use \( \theta \), and if not stated differently \( \theta \) is defined as \( \theta := \logit(p) = \log \left( \frac{p}{1-p} \right) \). In general \( \theta \) can also represent a continuous parameter on the range of real numbers. Index values 0 and 1 represent the null and alternative value of \( p \) and \( \theta \). In general \( d_{ij} \) stands for the pairwise distance measure between two baskets \( i \) and \( j \), specifications about the type of distance (or divergence) measure will be given in the respective sections. The variable \( \lambda \) describes the probability of a distribution in a Bayesian mixture distribution and the variable \( \omega \) represents a weight. A general parameter of a distribution is given by \( \gamma \). Thresholds for decision-making are denoted by \( \tau \), where index \( F \) indicates interim futility assessment, \( E \) interim efficacy assessment and \( A \) the analysis at end of the trial. Additional indices in superscript distinguish parameters of the same type but with different values. The function \( f() \) represents the density of a distribution. Distributions are given by \( \mathcal{G}, \mathcal{F} \) and \( \mathcal{E} \). In case of different distributions behind \( \mathcal{G} \) (or \( \mathcal{F}, \mathcal{E} \)) within one model, we apply indices to distinguish the distributions from each other. We deliberately use \( \mathcal{G}, \mathcal{F} \) and \( \mathcal{E} \) instead of explicit distributions to better focus on the technical ideas rather than on chosen distributions which can be changed and tuned for the respective trial purposes.

3.1 Component 1: Information sharing

The techniques for information sharing depend on the applied statistical approaches in the basket trial (frequentist, Bayesian). The techniques for information sharing are therefore presented according to the applied technique. The order of presentation is given by the evolution of techniques, their increasing complexity and their differing way of addressing the primary objective of the trial. Figure A2 displays the connections, how sharing techniques evolved from each other, and a categorization of the sharing techniques (see Appendix).

3.1.1 Frequentist: Pool all or nothing

The frequentist approaches use a pool all or nothing approach. The so far available data is evaluated at a predefined interim node of the trial and based on the results a sharing decision is taken. The two distinct decision options are to combine all baskets into one pooled data set or to consider all basket independent of each other with individual analyses and decisions in each basket. The proposed decision rules to proceed with either of the two options in the frequentist pool all or nothing approach are:

(i) Pass 1st stage—Chen et al. (2016), Li et al. (2019), Zhou et al. (2019)

The simplest decision rule is a combination of futility assessment and implicit homogeneity evaluation. All baskets that pass the first stage without futility stop are considered homogeneous regarding efficacy and therefore pooled after the interim analysis into one group. The others are stopped and not further investigated.

(ii) (Fisher’s) Exact test for contingency table—Cunanan et al. (2017b)

An (Fisher’s) Exact Test is applied to investigate the hypothesis of heterogeneous baskets regarding the responses. If the null hypothesis (\( H_0 \): homogeneous response rates) is rejected, each basket is investigated individually, whereas otherwise all baskets are pooled into one group. The critical value of the test statistic (or its significance level) is a tuning parameter and is predefined.

3.1.2 Bayesian techniques

The Bayesian techniques conduct the sharing of information with different methods and underlying ideas. The so far proposed techniques can be subdivided into three groups. The first group is the Bayesian hierarchical model with
normal distribution for the transformed response rate \( \hat{\theta}_i \) and we denote it as BHM. The simple BHM is the basic version of Bayesian information sharing and the root for the evolution of the sharing techniques for basket trials. It was first proposed by Thall et al. (2003) for the analysis of phase II trials where multiple subtypes are treated and the treatment effect among the subtypes are assumed exchangeable and correlated, but without direct reference to basket trials. Berry et al. (2013) adapted this approach to the developing research area of basket trials. The developed BHMs are either mean- or variance-driven. The second group shares information directly within the distribution of the response rate \( p_i \) without transformation and the third group shares information based on the probability of the investigated hypothesis.

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

The BHM assumes that the (transformed) response rate of each basket comes from the same normal distribution. That assumes the baskets are exchangeable which means that the joint distribution of the response rates is permutation-invariant. The defining parameters of the normal distribution (\( \mu \) and \( \sigma^2 \)) are random variables itself with their own distributions \( F \) and \( G \). The parameters that describe those distributions \( (F \) and \( G \)) are called hyperparameters and form the second stage of the BHM. They are usually set to a specific value. However, it is possible to define them as random variables as well, which forms another stage in the hierarchy. The BHM of Thall et al. (2003) and Berry et al. (2013) uses a two-stage hierarchical model, where \( \mu \) influences the location of the response rates and \( \sigma \) the amount of sharing between the baskets. For higher values of \( \sigma \), less information is shared and vice versa, low values of \( \sigma \) indicate a higher degree of sharing. For ease of clear and understandable presentation the explicit distributions and hyperparameters of the second stage are not shown, but included in \( F \) and \( G \). The distribution of \( \hat{\theta}_i \) generated from the BHM is therefore given by

\[
\begin{align*}
\hat{\theta}_i &\sim N(\mu, \sigma^2) \\
\mu &\sim G \quad \forall i = 1, \ldots, I. \\
\sigma^2 &\sim F
\end{align*}
\]

Note that Berry et al. (2013) transform the response rate to \( \hat{\theta}_i := \logit(p_i) - c \), where \( c \) is a constant reference value, which can be incorporated in the expected value \( \mu \) to work with \( \hat{\theta}_i := \logit(p_i) \).

**Mean-driven BHMs**

The basic BHM is extended by additional BHMs or individual distributions to cover a broader field of possible distributions including for example a distinct separation of a favorable (null) and non-favorable (alternative) scenario. The scenarios display the difference by a shifted expected value and the individual BHMs are applied with probability \( \lambda \).

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

The ExNex (exchangeable - nonexchangeable) design extends the BHM by an individual distribution. The individual distribution represents the case when the response rate does not fit with the ones in the other baskets and non-exchangeability between the baskets is assumed. Hence, the distribution of \( \hat{\theta}_i \) is given by

\[
\hat{\theta}_i \sim \begin{cases} 
N(\mu, \sigma^2) & \text{with probability } \lambda_i \\
\mu \sim G \\
\sigma^2 \sim F \\
N( m_i, v_i) & \text{with probability } (1 - \lambda_i) \\
m_i, v_i \text{ fixed} \\
\lambda_i \text{ fixed}
\end{cases} \quad \forall i = 1, \ldots, I.
\]

The parameters of the individual distribution are fixed and predefined values. The probabilities \( \lambda_i \) are fixed \textit{a priori} and different values for different baskets are possible. This could be the case when one \textit{a priori} assumes that basket \( i \) does not conform with the others or vice versa. Moreover Neuenschwander et al. (2016) mention
to include another hierarchical stage by defining $\lambda_i$ as a random variable with, for example, Dirichlet distributions.

(v) **PoC—Jin et al. (2020a)**

The next evolution step of information sharing is the mixture of two BHMs and is proposed in the proof of concept (PoC) design by Jin et al. (2020a). The two BHMs differ in the expected value around which they are centered. One BHM could represent a null scenario with a response rate that is considered clinically futile whereas the other BHM could represent an alternative value that indicates a promising treatment effect. The amount of sharing is given by the variance which is drawn from the same distribution in both BHMs. The probability $\lambda$ with which the BHMs are chosen is random and the same for all baskets. Hence, the distribution of $\theta_i$ is given by

$$
\theta_i \sim \begin{cases} 
N(\mu_1,\sigma^2_1) & \text{with probability } \lambda \\
\mu_1 \sim G_1 \\
\sigma^2_1 \sim P \\
N(\mu_2,\sigma^2_2) & \text{with probability } (1-\lambda) \\
\mu_2 \sim G_2 \\
\sigma^2_2 \sim P \\
\lambda \sim U[0,1] 
\end{cases} 
\quad \forall i = 1, \ldots, I.
$$

PoC applies an additional hierarchy stage for the parameters of $G_1$ and $G_2$.

(vi) **QBHM—Liu et al. (2017)**

The sharing technique proposed for the final analysis in QBHM is very similar to PoC. The differences are individual variances in the BHMs and a fixed probability for each BHM that is not further specified in the original paper. Before the mixture of BHMs is applied, a Cochran’s Q test is conducted and only if homogeneity was not rejected, the design continues with information sharing in the final analysis. Otherwise each basket is investigated independently. QBHM applies a modified test-then-pool approach where a test decides whether pooling is conducted, and if so, partial pooling is performed. We denote the design of Liu et al. (2017) as QBHM because of the Cochran’s Q test. The distribution of $\theta_i$ is given by

$$
\theta_i \sim \begin{cases} 
N(\mu_1,\sigma^2_1) & \text{with probability } \lambda \\
\mu_1 \sim G_1 \\
\sigma^2_1 \sim P \\
N(\mu_2,\sigma^2_2) & \text{with probability } (1-\lambda) \\
\mu_2 \sim G_2 \\
\sigma^2_2 \sim P \\
\lambda & \text{fixed} 
\end{cases} 
\quad \forall i = 1, \ldots, I.
$$

(vii) **BaCIS—Chen and Lee (2019)**

In the first step for the information sharing in the Bayesian classification and information sharing design (BaCIS) each basket is allocated to one of two clusters ($C_1$ or $C_2$). The clustering is based on two BHMs with different means where each represents one cluster. In the clustering process a latent variable assigns each basket to either $C_1$ or $C_2$. The final decision about cluster membership of each basket is made on the distribution of the latent variable. If the latent variable exceeds a reference value with a given probability, the basket is assigned to $C_1$, or otherwise, to $C_2$. For the probability threshold of 0.5, the unique cluster membership is determined by the higher frequency of assignment to each of the two BHMs. In the second step a BHM is executed within each cluster to share information only among the baskets from the same cluster. Consequently, the BaCIS approach represents as binary mixture of
two BHMs and the distribution of \( \theta_i \) is given by

\[
\theta_i \sim \begin{cases} 
N(\mu_1, \sigma^2_1)|D_{c_1} & \text{if basket } i \text{ located to cluster } C_1 \\
\mu_1 \sim \mathcal{G}_1 \\
\sigma^2_1 \sim F \\
N(\mu_2, \sigma^2_2)|D_{c_2} & \text{if basket } i \text{ located to cluster } C_2 \\
\mu_2 \sim \mathcal{G}_2 \\
\sigma^2_2 \sim F \\
\end{cases} \quad \forall i = 1, \ldots, I.
\]

(viii) **BLAST**—Chu and Yuan (2018b)

The Bayesian latent subgroup trial (BLAST) design shares information by a mixture of \( k \) BHMs. Each BHM represents a cluster of baskets with similar treatment effects. The latent cluster membership of each basket \( i \) is mainly based on the longitudinal trajectory of a biomarker which forms as a proxy for the binary response and on the observed responses. The number of subgroups \( k \) is defined by the goodness of fit in the semiparametric mixed model for the longitudinal trajectory. Chu and Yuan (2018b) justify the incorporation of the longitudinal trajectory by the limited success of clustering of responses in small sample sizes scenarios and by the limited information in the binary outcome. Hence, the distribution of \( \theta_i \) is given by

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

The \( k \)-dimensional vector of probabilities \( \lambda_i \) for each of the \( k \) BHMs follows a Dirichlet distribution.

**Variance-driven BHM**

The sharing in the BHM can be guided by the mean which imposes the common position, and by the variance which regulates the amount of information shared between baskets. With larger variance the position loses impact and a wide range of values can be taken and therefore less sharing takes place. On the other hand, a small variance does not allow many different values and therefore a lot of information is shared and the basket-individual parameters are all very similar. The extremes are \( \sigma^2 = \infty \) for complete independence and \( \sigma^2 = 0 \) for complete sharing. Proposed techniques that modify the variance to share information among baskets are presented here.

(ix) **calBHM**—Chu and Yuan (2018a)

The calibrated BHM (calBHM) adapts the basic BHM (cf. Section 3.1) by a fixed value for the variance. The new value for the variance is determined by a monotonically increasing function of the test statistic \( T \) from the \( \chi^2 \)-test and the tuning parameters \( \phi \). The function \( g : R^+ \to R^+ \) is calibrated with regard to \( \phi \) via simulations before the start of the trial to control the type 1 error. The authors allow in general any function that fulfills the monotonicity in \( T \) and present as an example \( g(T; \phi = (\phi_1, \phi_2)) := \exp(\phi_1 + \phi_2 \cdot \log(T)) \) with \( \phi_2 > 0 \) which showed robust characteristics in their simulations. The distribution of \( \theta_i \) is defined by

\[
\theta_i \sim N(\mu, \sigma^2) \\
\mu \sim \mathcal{G} \\
\sigma^2 = g(T; \phi) \
\]
The sharing technique of corBHM is given by a multivariate normal distribution with a correlation matrix $\Omega$ where each element $\omega_{ij}$ is generated by a correlation function of the pairwise distance measures $d_{ij}$ and the random tuning parameter $\phi \sim \mathcal{E}$. The distance is calculated on the posterior distributions of each individual basket. The authors propose the Kullback-Leibler distance, the Hellinger distance or the Bhattacharyya distance as the distance measure. The vector of basket-individual response rates is distributed as follows

$$\theta \sim MVN(\mu, \Omega \cdot \sigma^2_b + \text{diag}(\sigma^2_w))$$

$$\mu \sim \mathcal{G}$$

$$\sigma^2_b \sim \mathcal{F}_b, \quad \sigma^2_w \sim \mathcal{F}_w$$

The variable $\sigma^2_b$ describes the variance between the baskets and $\sigma^2_w$ the variance within a basket. The latter are stored as a diagonal matrix denoted by $\text{diag}(\sigma^2_w)$. corBHM applies an additional hierarchy stage for the variance in $\mathcal{G}$.

(xi) BCHM—Chen and Lee (2020)

The BCHM design shares information in a two-step procedure. In the first step, a clustering with a Dirichlet process mixture model (DPM) is conducted and the pairwise membership of baskets in the same cluster throughout the sampling process of the DPM is evaluated. This relative frequency describes the similarity of two baskets and is denoted by $\omega_{ij}$, which is increasing with stronger similarity and has its minimum at 0 and its maximum at 1. Hence, for $i = j$ it follows that $\omega_{ij} = 1$. The relative frequency $\omega_{ij}$ modifies the variance in the BHM in the second step and therefore determines the degree of sharing between the baskets. Consequently, each basket $j$ has an individual variance in the BHM with respect to basket $i$, which is given by the quotient of the variance of basket $i$, $\sigma_{ij}^2$, and the similarity measure $\omega_{ij}$. The reference to $i$ is denoted by the superscripted index in round brackets. The distribution of $\theta_i$ is then given by the $i$-th element from the BHM with respect to $i$ and can be denoted as

$$\theta_{ij}^{(i)} \sim N\left(\frac{\mu^{(i)}}{\omega_{ij}}, \frac{\sigma_{ij}^2}{\omega_{ij}}\right) \quad \forall i = 1, \ldots, I$$

$$\mu^{(i)} \sim \mathcal{G}$$

$$\sigma_{ij}^2 \sim \mathcal{F}$$

$$\theta_i := \theta_{ij}^{(i)} \quad \forall i = 1, \ldots, I.$$

Thus, higher similarity leads to a smaller variance and therefore to more sharing between $j$ and $i$. Vice versa, for lower similarity, the variance increases which then reduces the amount of sharing between the baskets.

(xii) Zheng—Zheng and Wason (2019)

Zheng and Wason (2019) share information between baskets with a marginal predictive prior (MPP). The MPP is created for each basket $i$ and is a weighted linear combination of $(I - 1)$ random variables. Each element is a commensurate predictive prior (CPP$_{ij}$) for basket $i$ based on basket $j$ with $j \neq i$. The weights $\omega_{ij}$ are functions of the Hellinger distance between basket $i$ and $j$. The CPP$_{ij}$ for $i$ based on $j$ is created from a normal distribution. The mean $\mu_j$ is fixed and inferred from the observed results in basket $j$. The random variance has a distribution which depends on the similarity between the results in basket $i$ and $j$ and therefore controls the amount of shared information from $j$ to $i$. For the CPP$_{ij}$ the variance $\sigma_{ij}^2$ is integrated out. For the analysis of basket $i$, the MPP is updated by the observed data in basket $i$. Hence, to get the distribution of $\theta_i$ the following steps are conducted

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

$\mu_j$ 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 $\sigma_{ij}^2$ which results in

$$\theta_{ij}^{(i)} \sim N(\mu_j, \sigma_{ij}^2)$$
with inferred and non-random values for \( \mu_j, \sigma^2_j \)

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

\[
\theta_i := \sum_{j \neq i} \omega_{ij} \cdot \theta^{(i)}_j \sim N(\sum_{j \neq i} \omega_{ij} \cdot m_j, \sum_{j \neq i} \omega^2_{ij} \cdot \sigma^2_j) \quad \forall i = 1, \ldots, I.
\]

Non-transformed and beta-binomial model
The so far presented Bayesian sharing techniques all work with the transformed response rate \( \theta_i \). However, there are techniques that enable sharing of information directly on \( p_i \). Those techniques take advantage of the conjugate property of binomial data (\( r \) responses among \( n \) observations) in combination with beta distributed response rates which results in a beta-binomial model.

(xiii) Simon—Simon et al. (2016)

In Simon’s basket trial design the response rate is categorical \( p \in \{p_0, p_1\} \) and \( p_i \) takes the value \( p_0 \) with probability \( \gamma \). This categorical distribution is denoted by \( \text{Cat}(p_0, p_1; \gamma) \). The information sharing by Simon assumes two scenarios. The homogeneous scenario accounts for exchangeability of the response rate among all baskets and evaluates all baskets together and equally. The second scenario assumes independence between the baskets and evaluates each basket individually. These assumptions are similar to the ExNex approach whereas in Simon’s design the response rate is categorical and the data are fully shared while ExNex allows an adjusted degree of sharing in the BHM. The separation into two terms with complete sharing and individual evaluation reflects a Bayesian all or nothing approach. Consequently, the distribution of the basket-individual response rate is given by

\[
p_i \sim \begin{cases} 
\text{Cat}(p_0, p_1; \gamma) | D_i & \text{with probability } \lambda \\
\text{Cat}(p_0, p_1; \gamma) | D_i & \text{with probability } (1 - \lambda) 
\end{cases} \quad \forall i = 1, \ldots, I.
\]

The parameter \( \lambda \) is fixed and the same for all baskets, which is not further specified in the original paper. The posterior probability of \( p_i \) can be calculated in closed form with the Bayes theorem, first published by Bayes and Price (1763).

(xiv) Asano—Asano and Hirakawa (2020)

The proposed sharing technique is a combination of Simon et al. (2016) with continuous \( p \) and is originally written as a Bayesian model averaging approach as proposed by Psioda et al. (2019). The sharing technique reflects a Bayesian all or nothing approach. All information is pooled among all baskets in the homogeneous scenario \((p_1 = \cdots = p_I)\). The fixed parameter \( \lambda \) describes the probability for the homogeneous scenario, \((1 - \lambda)\) the probability for the heterogeneous scenario, respectively. In the heterogeneous scenario each basket is considered individually; however, two different priors are used, thus allowing to distinguish between different scenarios, for example, effective and ineffective responses based on already available information. The fixed probability for the first prior is \( \omega \) and \((1 - \omega)\) for the second. Hence, Asano and Hirakawa (2020) define the distribution of the response rate as

\[
p_i \sim \begin{cases} 
\text{Beta} \left( a + \sum_{j} r_j, b + \sum_{j} (n_j - r_j) \right) & \text{with probability } \lambda \\
\text{Beta} \left( a_1 + r_i, b_1 + (n_i - r_i) \right) & \text{with probability } (1 - \lambda) \cdot \omega \\
\text{Beta} \left( a_2 + r_i, b_2 + (n_i - r_i) \right) & \text{with probability } (1 - \lambda) \cdot (1 - \omega)
\end{cases} \quad \forall i = 1, \ldots, I.
\]

(xv) MEM—Hobbs and Landin (2018)

The sharing technique in the multi-exchangeability model (MEM) is based on the exchangeability matrix \( \Omega \) which displays in a pairwise manner if two baskets are either considered exchangeable \((\omega_{ij} = 1)\) or not \((\omega_{ij} = 0)\). The matrix \( \Omega \) is random with for example, \( a \text{ priori} \) equal probabilities \( P[\omega_{ij} = 1] = 0.5 \). If \( \Omega \) indicates exchangeability between two baskets, the observed results in both baskets are combined, otherwise they remain separated. Finally, all possible exchangeable-nonexchangeable combinations for all \( I \) baskets are modeled in combination with the observed data. The resulting beta-binomial model returns the posterior distributions of the basket-wise response
rates in the following form
\[ p_i \sim \text{Beta}\left(a + \sum_j \omega_{ij} \cdot r_j, b + \sum_j \omega_{ij} \cdot (n_j - r_j)\right) \quad \forall i = 1, \ldots, I. \]
\( a, b \) fixed values
\[ \Omega \sim \mathcal{E} \]

(xvi) **BMA**—Psioda et al. (2019)
The Bayesian model averaging (BMA) approach applies the same idea as Hobbs and Landin (2018) with their **MEM** approach. All possible models, each presenting a different situation of exchangeability among all baskets, are evaluated together with the observed data and the posterior distributions of the basket-wise response rates are calculated from a beta-binomial model.

(xvii) **Fujikawa**—Fujikawa et al. (2020)
Fujikawa uses the properties of the beta-binomial model to incorporate information from other baskets. The information from basket \( j \) is shared with basket \( i \) by incorporating the individual posterior distribution in weighted form. The weight \( \omega_{ij} \) is a function of the Jensen-Shannon divergence of the individual posterior distributions from basket \( i \) and \( j \), and two fixed tuning parameters. The basket-specific priors are defined by \( a_j, b_j \). Hence, the basket-individual response rate is distributed as follows
\[ p_i \sim \text{Beta}\left(\sum_j \omega_{ij} \cdot (a_j + r_j), \sum_j \omega_{ij} \cdot (b_j + n_j - r_j)\right) \quad \forall i = 1, \ldots, I \]
\( a_j, b_j \) fixed values

**Hypothesis-driven models**
The hypothesis-driven sharing techniques concentrate on the probabilities for the null and alternative hypothesis. Information is shared among baskets with similar probabilities for the hypothesis and not based on similar response rates.

(xviii) **RoBoT**—Zhou and Ji (2020)
The sharing of information in the robust Bayesian hypothesis testing method (RoBoT) is conducted within latent subgroups. The latent subgroups are determined by a Dirichlet process mixture model (DPM) based on the probabilities of basket-individual hypotheses \( H_0 : p_i \leq p_{i0}, H_1 : p_i > p_{i0} \). The Dirichlet process DP is defined by the fixed concentration parameter \( \alpha \) and the given base distribution \( G_0 \). Hence, the distribution of the response rate \( p_i \) is given by
\[ p_i \sim \begin{cases} 
\text{TBeta}(a_0, b_0, [0, p_{i0}]), & \text{with probability } \lambda_i \\
\text{TBeta}(a_1, b_1, (p_{i0}, 1]), & \text{with probability } (1 - \lambda_i) 
\end{cases} \quad \forall i = 1, \ldots, I \]
\( a_0, b_0 \) fixed
\( a_1, b_1 \) fixed
\[ \text{logit} (\lambda_i) \sim N(\mu_i, \sigma_i^2) \]
\( (\mu_i, \sigma_i^2) \sim G \]
\( G \sim \text{DP}(\alpha, G_0) \),

where \( \text{TBeta}(a, b, I) \) stands for a truncated beta distribution with parameters \( a, b \) and restricted support to the interval \( I \).

(xix) **MUCE**—Lyu et al. (2020)
The information sharing in the MUCE design consists of a hierarchical model for the basket-individual hypotheses. The null and alternative hypothesis are indicated by the random variable $Z_i$ which is shaped in a hierarchical model to share information among different baskets. Consequently, $\theta_i$ is denoted by

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

$$Z_i \sim N(\mu_i, \sigma_i^2), \quad \sigma_i^2 \text{ fixed}$$

$$\mu_i \sim N(\mu_0, \sigma_0^2), \quad \sigma_0^2 \text{ fixed}$$

$$\mu_0 \sim N(\mu, \sigma^2), \quad \sigma \text{ fixed},$$

where $TCauchy(\theta, \gamma, I)$ stands for a truncated Cauchy distribution with location parameter $\theta$, scale parameter $\gamma$, and restricted support to the interval $I$.

### 3.2 Component 2: Interim futility assessment

In this section, we present the different tools that are proposed for the interim futility assessment. The futility tools define the rules that indicate whether a basket is pruned or continued at the interim assessment. Moreover, the technical relationships between the techniques are drawn. The futility decisions are made directly on the response rate $p_i$, designs that work with transformations rescale $\theta_i$ to $p_i$. Information sharing can be conducted before a futility assessment. This is a design aspect defined by the alignment of the components in the underlying trial.

(i) **Minimum number of responses**: prune basket $i$ if fewer 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 (regarding the primary endpoint and $H_0 : p \leq p_0, H_1 : p > p_0$) exceeds $\alpha_F$. The significance level is chosen in the context of the whole trial design, but is supposed to be higher than the commonly used 5% for statistical tests, because otherwise the majority of the baskets would be pruned.

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.,

$$P[p_i > p_0 | D] \leq \tau_F.$$  \hspace{1cm} (1)

Used by: Hobbs and Landin (2018), Thall et al. (2003), Asano and Hirakawa (2020).

One modification is to change the reference value to $\frac{p_0 + p_1}{2}$. This fraction includes the alternative value for the response rate and causes a shift in the reference value, that is,

$$P[p_i > \frac{p_0 + p_1}{2} | D] \leq \tau_F.$$  \hspace{1cm} (2)

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

The second modification for the posterior probability reflects the assumed discrete distribution of the response rate which can only take values $p_0$ or $p_1$, that is,

$$P[p_i = p_1 | D] \leq \tau_F.$$  \hspace{1cm} (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. Lyu et al. (2020) uses the random variable \( Z_i \) to indicate the distribution of the response rate in basket \( i \). If \( Z_i \geq 0 \), the basket follows the distribution given by the alternative hypothesis, otherwise the distribution is given according to the null hypothesis. Hence, basket \( i \) is pruned if

\[
P[Z_i \geq 0 | D] < \tau_F.
\]

Used by: Lyu et al. (2020).

(v) **Posterior predictive probability**: prune basket \( i \) if the predictive probability that the response rate exceeds the reference value at the final analysis is low. The final number of observations per basket must be known when the Bayesian predictive probability is applied.

Liu et al. (2017) simulate the number of responses for the second stage based on the so far observed data. At interim, they draw stage two responses from a beta-binomial model that was generated by the stage one data. With the drawn responses they calculate the final analysis. The posterior predictive probability is then estimated by the frequency of how often the final response rate is larger than the reference response rate \( p_0 \). The basket is pruned if the estimated posterior predictive probability is lower than the predefined threshold \( \tau_F \).

Fujikawa et al. (2020) also use a beta-binomial model for the response rate but compute the posterior predictive probability analytically and prune a basket if the following equation holds true

\[
\sum_{\{r_i: \text{final analysis successful}\}} \int_0^1 f(r_i | p_i) f(p_i | D) dp_i \leq \tau_F.
\]

(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. The number of remaining observations must be known and an assumption about the response rate must be made, resulting in

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

(5)

The assumed response rate \( \hat{p} \) can be the assumed rate under the alternative hypothesis or could be given by the maximum likelihood estimate (MLE) of the response rate based on the observed data at interim (Saville et al., 2014). Other options for the point estimate of the assumed response rate could be boundaries of (e.g., 90%) confidence intervals. They would incorporate at least partly the uncertainty regarding the true response rate as this is ignored by the point estimate using the ML-method.

Used by: so far not applied in proposed basket designs but a suitable frequentist tool.

For Zheng and Wason (2019), Li et al. (2019), and Chen and Lee (2019) interim futility assessments are optional, without further explicit explanation. In those cases the technique of information sharing is the core element of their publication and not the complete design. Neuenschwander et al. (2016), Jin et al. (2020a), Chen and Lee (2020), Zhou and Ji (2020) do not mention interim futility assessments at all. Nevertheless, their techniques can be applied in a trial design with futility assessments.

Each futility tool has a different way to process the available information to decide whether to prune a basket or not. The futility tools are connected with each other, which we present in the following. The connections underline the flexibility of each technique as part of a modular setup. Moreover, it should be noted that the tools can be tuned to come to the same conclusions at the same decision node of a trial. The interaction between the futility tools is displayed in Figure A3 together with the information in which design they are applied.

Let \( D = \{n, r\} \) be the available observed information at the interim futility assessment in basket \( i \). The index is left out for purpose of easier readability. The minimum response criterion is fulfilled if the number of observed responses is \( r \geq r_{\text{crit}} \).

At the same time a statistical test with \( H_0 : p \leq p_0 \) and \( H_1 : p > p_0 \) returns a certain p-value for \( r_{\text{crit}} \) responses. If this p-value is smaller than the significance level \( \alpha_F \) the null hypothesis is rejected, which indicates that currently there is enough support for a non-futile effect. Hence, at a known number of observations \( n \) the significance level \( \alpha_F \) can be chosen that for at least \( r_{\text{crit}} \) responses the statistical test does not stop for futility. Moreover, a minimum number of responses can be
considered as a primitive one-sided binomial test. Vice versa a minimum number of responses \( r \) among \( n \) observations is equivalent with a certain significance level \( \alpha_F \) (e.g., for \( n_1 = 15, p_0 = 0.3, \) and \( \alpha_F = 0.25 \) or for at least \( r = 7 \) responses, the null hypothesis \( H_0 \) is rejected in a one-sided binomial test).

We now know that there is a smallest \( r_{crit} \) such that the p-value is smaller than \( \alpha_F \) and the basket is not pruned. Consequently, in a Bayesian setting, we can choose the threshold \( \tau_F \) for the posterior distribution such that it fulfills \( P[p > p_0|n, r_{crit}] > \tau_f \) and that for \( r < r_{crit} \) the posterior probability is smaller than or equal to the \( \tau_F \) (cf. Equation 1). However, this connection between the frequentist statistical test and the Bayesian posterior probability only holds that easily if no information sharing was conducted in advance. The reason is that with shared information the threshold \( \tau_F \) not only depends on \( n \) and \( r \) of the current basket but also on the observations in the remaining baskets and on the sharing technique with its parameters.

The futility assessment with the posterior probability offers two reference values, \( p_0 \) (Equation 1) or \( \frac{p_0+p_1}{2} \) (Equation 2). Both reference values use the same posterior distribution of response rate \( p \). The difference between the two tools lies in the investigated intervals on the space of \( p \) due to the different reference values. Without loss of generality, we assume \( p_0 < p_1 \). For \( r_{PP} \) the posterior probability indicates no futility stop because \( P[p > \frac{p_0+p_1}{2} | n, r_{PP}] > \tau_m \) and for \( r < r_{PP} \) the basket is pruned. Now the question is how to choose \( \tau_F \) to come to the same conclusion for reference value \( p_0 \). We know that

\[
P[p > \frac{p_0 + p_1}{2} | n, r] = P[p > p_0 | n, r] - P[p \in \left( p_0, \frac{p_0 + p_1}{2} \right) | n, r] \leq \tau_m^F
\]

and to come to the same conclusion we require at the same time \( P[p > p_0 | n, r] \leq \tau_F \). Therefore it follows from

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

that \( \tau_F = \tau_m^F + P[p \in (p_0, \frac{p_0 + p_1}{2}) | n, r] \) with the last summand being a constant value. To assure consistency, the constant increment to \( \tau_m^F \) is the probability mass of the investigated posterior distribution in \( (p_0, \frac{p_0 + p_1}{2}) \). Hence, the same decisions in both tools are taken if reference values and thresholds are adapted. Adaptations in the reference value and the thresholds are interchangeable because both futility tools rely on the same posterior distribution which can include shared information from other baskets.

The design proposed by Lyu et al. (2020) uses a hypothesis-driven approach where the posterior distribution of the transformed response rate \( \theta \) is separated by truncated Cauchy distributions at the null reference \( \theta_0 \) (cf. Section 3.1). A basket is pruned for futility if \( P[Z \geq 0|D] < \tau_F \), and \( Z \geq 0 \) represents the case when \( \theta \) is located below \( \theta_0 \). Therefore, the futility tool applied in Lyu et al. (2020) can be rewritten as \( P[p > p_0|D] < \tau_F \) and conforms with the posterior probability tool.

The posterior predictive probability (PPP) is a function of the current posterior distribution and the number of remaining observations \( n_2 \) (cf. Equation 4). The PPP calculates the total probability for a successful final analysis based on all information so far available. The PPP tool uses the same threshold \( \tau_F^P \) for every reference value and every current number of observations. This is not the case for the posterior probability because for smaller \( n_1 \) the variance of the posterior distribution is larger which means the certainty is lower and therefore the threshold \( \tau_F \) must be adapted according to \( n_1 \). Nevertheless both tools go hand in hand as the threshold \( \tau_F \), given \( n_1 \) observations, can be adapted such that it corresponds to the same decision as PPP. There is a smallest number of responses \( r_{PPP} \) for which PPP continues and hence for \( \tau_F := P[p > p_0 | n_1, r_{PPP} - 1] \), the posterior distribution induces the same decision. The PPP is either simulated or directly calculated. The simulation converges towards the calculated PPP due to the law of large numbers. However, it might cost some numerical computation time. The direct calculation works in Fujikawa et al. (2020) because a beta-binomial model is applied, hence the posterior distribution is beta-distributed and the number of future responses follows a binomial distribution which finally results in a sum over beta functions (see Appendix A.1).

The conditional power is a function of the assumed response rate, the number of additional observations, and the significance level \( \alpha \) for the final analysis. If the assumed response rate reflects the so far observed responses \( r \), the threshold \( \tau_F \) can be tuned in accordance with the minimum number of responses \( r_{crit} \) to not prune that basket. The posterior power tool is then in line with the statistical test and the minimum number of responses. The conditional power and the posterior predictive probability pursue the same idea. Both look ahead to the final analysis and evaluate if it is worth to continue, and otherwise stop. Moreover, they show the difference between the frequentist and the Bayesian approach because the
conditional power (cf. Equation 5) uses a point estimate for the assumed response rate, whereas the posterior predictive probability (cf. Equation 4) considers the response rate as a random variable with respective (posterior) distribution.

We have seen that the futility tools are connected with each other. However, the number of observations \( n \) at an interim futility assessment must be known to tune the parameters and thresholds such that the same decisions are drawn. Moreover, for the posterior predictive probability (and the conditional power) the final number of observations must be predefined. Information sharing before futility assessments cuts the direct and easy connection between the frequentist and the Bayesian approaches. This is due to the fact that after sharing the decisions in every basket also depend on the results in the other baskets and on the sharing technique (with its potentially numerous parameters).

The frequency of interim futility assessments is a design element of the trial. The futility assessments are optional, so they do not have to be included, but when they are, their time points and rules must be prespecified. Basket trial designs with one interim futility assessment are proposed in (Chen et al., 2016; Cunanan et al., 2017a; Jin et al., 2020b; Liu et al., 2017; Zhou et al., 2019). Several interim futility assessments are allowed in (Psioda et al., 2019; Chu & Yuan, 2018b; Chu & Yuan, 2018a; Fujikawa et al., 2020; Li et al., 2019; Hobbs & Landin, 2018; Lyu et al., 2020; Simon et al., 2016). The assessments take place in dependence of the passed study time or on the observed data. Berry et al. (2013) propose to conduct the futility assessments after a certain number (e.g., 10) of patients per basket have been observed, further interim looks take place when additional patients (e.g., 5) were observed in each basket. In the most extreme scenario, futility assessments are be conducted after each observation (Simon et al., 2016). On the other hand one single interim futility assessment (e.g. Jin et al., 2020b) results in a 2-stage design.

3.3 Component 3: Interim efficacy assessment

The interim efficacy assessment is an optional element which can be incorporated to stop baskets early due to strong evidence of a successful treatment. This component is explicitly proposed only by few designs (Chen et al., 2016; Psioda et al., 2019; Fujikawa et al., 2020; Simon et al., 2016) and mentioned as an optional element in (Berry et al., 2013; Chen & Lee, 2019; Zheng & Wason, 2019) without further technical definitions in the former two. All other designs do not consider interim efficacy assessments. The proposed tools for the interim efficacy assessments are:

(i) **Statistical test**: prune basket \( i \) if the p-value from an appropriate statistical test (regarding the endpoint) exceeds \( \alpha_E \). The significance level is chosen in the context of the whole trial design and in accordance with the regulatory requirements for accelerated approval based on a confirmatory phase III basket trial.

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.

\[
P[p_i > p_0 | D] > \tau_E. \tag{6}
\]

Used by: Psioda et al. (2019).

As for the futility assessment one can change the reference value which then results in

\[
P \left[ p_i > \frac{p_0 + p_1}{2} | D \right] > \tau_E. \tag{7}
\]

Used by: Berry et al. (2013).

In the case of a discrete space for the response rate, the posterior probability is adapted, respectively,

\[
P[p_i = p_1 | D] > \tau_E. \tag{8}
\]

Used by: Simon et al. (2016).
(iii) **Posterior predictive probability**: prune basket $i$ if the predictive probability that the response rate exceeds the reference value at the final analysis is high, i.e.,

$$\sum_{\{r_i: \text{final analysis successful}\}} \int_0^1 f(r_i|p_i)f(p_i|D)dp_i > \tau_E.$$  \hfill (9)

Used by: Fujikawa et al. (2020).

The applied techniques are very similar to the futility tools. They only adapt the direction of the predefined threshold from $<$ to $>$ (cf. Equations 2 and 7, Equations 1 and 6, Equations 3 and 8, and Equations 4 and 9). Consequently, the connections between the futility tools also apply for the interim efficacy tools. Psioda et al. (2019) use different reference values in the futility and efficacy assessments while the other designs remain consistent in the applied tool and only change the direction of the threshold. As for the futility assessments, the time points for the interim efficacy assessments must be prespecified. One pragmatic approach would be the simultaneous assessment of interim efficacy and futility at the same node of the trial. But it is also possible to apply the efficacy assessment independently at nodes where they pose a meaningful tool for the purpose of the trial. The necessity of an interim efficacy assessment must be seen in the context of the trial purpose. The focus in early trials is rather on pruning baskets with no effect and continuation with potentially promising indications than early detection of efficacy. Still, it leaves the door open for a quick development of a potential breakthrough treatment with overwhelming early results.

### 3.4 Component 4: Final analysis

The final analysis determines the success of a basket trial. The observed results in each basket and throughout the complete trial are evaluated with respect to the underlying research question (confirmatory evaluation of the investigated treatment, detection of promising baskets, PoC). Proposed tools to conduct the final analysis are:

(i) **Statistical test**: The frequentist designs apply a statistical test to assess the final efficacy.

The Binomial test is the first choice for the evaluation of a response rate and is conducted in a one-sided ($H_0 : p \leq p_0, H_0 : p > p_0$) or two-sided ($H_0 : p = p_0, H_0 : p \neq p_0$) fashion. Cunanan et al. (2017b) apply the one-sided version and Zhou et al. (2019) use the two-sided version in the final analysis of their exploratory trial design. 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 in line with the trial’s primary endpoint for the final analysis.

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

$$P[p_i > p_0|D] > \tau_A$$  \hfill (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).

There are modifications of the posterior probability decision rule similar to the ones for the interim futility and efficacy tools. One can additionally allow equality of the posterior distribution with the threshold $\tau_A$ to declare efficacy in the final analysis, that is,

$$P[p_i > p_0|D] \geq \tau_A.$$  

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 $\delta$ to incorporate an additional margin which must be exceeded to declare a basket successful, that is,

$$P[p_i > p_0 + \delta | D] > \tau_A.$$  

Used by: Zheng and Wason (2019), Chen and Lee (2020).

Equivalent to the interchangeable adaption of the reference value and the threshold in the futility tools, one can do so for the final analysis tools and abstain from $\delta$ and increase $\tau_A$ correspondingly (see Section 3.2). For a discrete distribution of the response rate, the posterior probability is adapted to assess the probability of the alternative hypothesis, that is,

$$P[p_i = p_1 | D] > \tau_A.$$  

Used by: Simon et al. (2016).

The posterior probability of the alternative hypothesis is applied in Lyu et al. (2020), since the random variable $Z_i$ uniquely indicates whether the response rate takes values larger or smaller than the reference value $p_0$. Consequently the following is equal to Equation 10

$$P[Z_i \geq 0 | D] > \tau_A.$$  

Used by: Lyu et al. (2020).

The included information $D$ in the final analysis of each basket depends on the sharing technique and therefore varies among the designs. Chen and Lee (2019) for example only use the information from the cluster the basket belongs to, while others, like Fujikawa et al. (2020), consider all available information.

(iii) **Proof of concept probability**: The PoC design evaluates in the final analysis if there is at least one basket in which the treatment shows efficacious results. A basket is considered efficacious if its posterior response rate is distributed according to the alternative hypothesis. This distribution is represented by one of the two BHMs (cf. sharing of PoC design), resulting in the criterion

$$P[\text{at least one efficacious basket} | D] = P\left[ \sum_{i=1}^{K} I_{\{\text{basket } i \text{ efficacious}\}} > 0 | D \right] > \tau_A.$$  

Used by: Jin et al. (2020a).

In Zhou and Ji (2020) no explicit tool for the final analysis of the RoBoT design is presented because the focus lies on the sharing technique. However, a final analysis can be conducted with a tool appropriate for the presentation of the response rate. An approach similar to Equation 11 could be applied, because the RoBoT design investigates the probabilities of hypothesis.

**4 PRACTICAL PROPOSAL FOR MODULAR CONSTRUCTION OF A BASKET TRIAL**

The previous section demonstrated that the tools for each component can be complicated and their arrangement can therefore result in a complex trial design with many parameters. Assumptions must be made for the (hyper-)parameters, thresholds, reference values and distributions. The parameter choices represent the current knowledge and situation in the investigated field of research or the parameters are tuned to ensure certain characteristics (TIE, power) of the trial design. The assumptions for the parameters apply to both the frequentist and the Bayesian designs and are a challenge in the practical application of basket trials. Moreover, the number of available and different tools for each component (sharing, interim futility, interim efficacy, and final assessment) is large and leaves at first sight an unclear situation. We propose to keep a basket trial design as simple as possible but as sophisticated as necessary to overcome researchers’ reservation in application of basket trials in clinical practice. Our proposed guidance to achieve this consists of the following aspects.
(i) Application of the modular set-up with four components and careful consideration whether the optional components introduce benefit with regard to the primary research question of the trial.
(ii) Transformation of the parameter of interest only if necessary (e.g., $\theta := \logit(p)$).
(iii) Consistent rules for interim futility, interim efficacy, and final assessment. The cut-off probability can change but the reference values should remain constant throughout the trial.
(iv) Tool for sharing of information is as understandable as possible. Sophisticated methods require justification why they should be preferred over less complicated methods. Direct comparisons of the most recent tools are so far not available.
(v) Number of parameters is kept as low as possible with focus on the important ones.
(vi) Thorough justification of the chosen prior values and distributions.
(vii) Proactive communication of statistical decisions and properties of the basket trial with all involved policymakers.

5 | DISCUSSION

Basket trials are a prevailing and dynamical field of research regarding methodological improvements and practical application in the development of innovative treatments (Meyer et al., 2020; Park et al., 2019). The progress in recent years has brought up many different trial designs and statistical methodologies. This has left an unclear situation in the literature about the existing methods and designs, which consequently impairs the practical application of basket trials. Hence, we created an overview with focus on the technical aspects of proposed methodologies and designs using a consistent notation. This facilitates the introduction to basket trial designs for statisticians and interested medical staff, it promotes the communication and planning of new basket trials in applied research, and it opens the door for further research on the statistical methods as differences, similarities, and opportunities are more easily detected. Moreover, this work resulted in a modular partition of basket trials into four components. The understanding of the applied techniques in each component enables a modular construction of basket trials consisting of known and future innovative techniques.

The categorization of basket trials into frequentist designs, Bayesian designs, and combination of both shows researchers the spectrum of statistical techniques and implicitly brings up the well known advantages and disadvantages of frequentist and Bayesian approaches. The second metric of categorization is the purpose of the trial throughout the phases of clinical development. The majority of designs are located in early phases, while only few concentrate on phase III trials. The late phases use frequentist approaches, which reflects the generally reluctant position of regulatory authorities regarding Bayesian approaches for approval of new treatments. Additionally, the late phases impose various challenges which impede the application of basket trials at this stage of drug development. The first challenge is the primary endpoint. While in phase II, a binary response might be sufficient to indicate an advantage of the treatment, phase III demands a stronger endpoint which without doubt reflects a benefit for the treated patients. In oncology, these late phase endpoints are usually time-to-event endpoints such as overall survival. In other indications, the endpoint could be a continuous variable measured at a given time point. Using a time-to-event endpoint increases the statistical complexity, as the interim futility and/or efficacy decisions must be considered more carefully because the primary endpoint might not be readily available at that node of the trial. An option to tackle this problem are surrogate endpoints, like for example progression-free-survival for the overall survival. However, the use of surrogate endpoints must be justified precisely since there needs to be clear evidence that the surrogate is capable of reliably predicting the primary endpoint (Fleming & LeMets, 1996). Apart from that, the sharing of information between the baskets bears more challenges. Appropriate statistical methods and tools are required if one aims to share information in a more regulated and sophisticated manner compared to a pool all or nothing approach. In that case, Bayesian techniques are of advantage, as we have seen by the variety of sharing techniques for binary outcomes presented in this manuscript. However, the complexity of the sharing tools might be accompanied with even more methodological challenges when the primary endpoint is a time-to-event endpoint. Another important challenge for a confirmatory setting is the required control group with randomized assignment of the patients. The control group must be basket-specific, unless there is a medical justification for a common control group for all baskets. This might impose logistical challenges because the number of patients can be restricted due to the rarity of the disease. Another issue is that the inference of a treatment effect in a specific basket/indication becomes challenging when information from other indications has been shared throughout the whole trial. The treatment effect must then be seen in the context of all baskets. Apart from the sharing, the pruning of baskets influences the control of errors. This characteristic is generally important in clinical trials and the confirmatory setting demands stricter control of errors than in
the exploratory setting. The challenges of error control are discussed later on in more detail. Regulatory expectations for confirmatory basket trials, including the control of multiplicity when information is shared, were recently presented by Collignon et al. (2020). The discussed aspects, as well as the more detailed discussion in Beckman et al. (2016), show that the hurdles for confirmatory trials are higher than for exploratory settings. However, further research could find solutions that facilitate confirmatory basket trials also with Bayesian methods. The recent publication by Lin et al. (2020), which is later discussed, could be a methodological starting point to fill this gap. We believe that the two metrics, statistical technique and purpose, currently constitute the most efficient way to categorize basket trials. However, we do not claim that this is the only way to classify basket designs nor that the classification is free from changes when new designs and techniques are published.

The sharing of information between baskets is the key component of a basket trial. It reflects the initial assumption that the treatment exploits the genetic predisposition of the patient (or his disease) irrespective of the disease location. The challenge of the trial designs and statistical techniques lies in an adequate procedure on how to share the information among the baskets. The most naive approach is to pool all baskets or to do nothing and evaluate each basket individually. This pool all or nothing approach only allows a two-way decision with either a homogeneous or heterogeneous path in the trial. The former is chosen if a minimum number of responses have been observed, or alternatively the decision is based on a statistical test. We see a statistical test rather critically as it leads to a separative decision without space for intermediate situations. Moreover, the direction of the hypothesis is wrong because the null hypothesis states homogeneous baskets, which actually is what one aims to show, and therefore one falsely remains with the null hypothesis due to low power. Both arguments stand in line with Neuenschwander et al. (2016)'s critique of limited benefit in early development where more tailored solutions are required to answer the underlying research question. The pool all or nothing approach is very prevalent in frequentist confirmatory trials where previous phases of development indicated efficacy. Liu et al. (2017) introduced a variation of the pool all or nothing approach. A Cochran's Q test decides on the homogeneous or heterogeneous path (with the same problem of wrongly directed hypotheses and lacking power) but the sharing is then conducted with a BHM which is not (necessarily) as strong in sharing of information as the pool all procedure. The advantages of the BHM are that the amount of sharing can be tuned through the parameters assigned to the prior distribution, and that it follows an intuitive idea that the response rates per basket all come from the same distribution. Therefore, the sharing is expressed by a regression to the mean such that the observed response rates are increased or decreased towards the overall mean (Berry, 2015). However, this idea assumes that all baskets are exchangeable which reflects the initial idea of similar behavior of all baskets. If this assumption turns out to be false, ineffective baskets are diluted by the shared information and wrongfully further investigated. This increases the type I error and ties resources in vain. Moreover, the BHM has problems with thorough sharing when there are only few baskets (Freidlin & Korn, 2013). To account for potentially nonexchangeable baskets, Neuenschwander et al. (2016) model the response rate as a mixture of two distributions, an exchangeable and a nonexchangeable part. This is a pragmatic way to deal with potentially violated assumptions and it initiated modifications of BHMs to improve their performance. Modifications concentrate on the different means in the distributions, compose the model of multiple elements, and adapt the variance to regulate the amount of sharing. The first two modifications offer a separation of the response space and therefore allow to share information among baskets with rather similar response to the treatment. This is applied in, for example, Jin et al. (2020a) which consists of two different BHMs where each can stand for a different clinical scenario, for example a promising response rate and a rather futile one. From our perspective, especially the modification of the variance is an appealing idea because the amount of sharing can be controlled and adapted individually, and consequently, a subset of baskets can share more information between each other (Chen & Lee, 2020; Jin et al., 2020b). The idea of partially exchangeable baskets originates in Hobbs and Landin (2018) as they implement a multisource exchangeability model (MEM) for a basket trial to leave behind the single source exchangeability models (SEM) like the BHM by (Berry et al., 2013; Thall et al., 2003). The multisource exchangeability model was initially developed by Kaizer et al. (2017) to incorporate external/supplemental data. For the sharing the MEM iterates over all possible pairwise exchangeability combinations among all I baskets. This implies a certain computational burden. Likewise, Psioda et al. (2019) propose the same appealing idea to go over the complete model space in their Bayesian model approach (BMA). Seeing two independent publications with the same idea underlines the quality of the approach. Unlike the majority of the other approaches, Hobbs and Landin (2018) and Psioda et al. (2019) directly model the response rate on its space [0,1] with a binomial distribution of the data and a beta prior which results in the conjugate Bayesian beta-binomial model. The advantages lie in the closed analytical form of the posterior distribution, an easy way to incorporate shared data by just adapting the parameters of the beta distribution by the number of responses and observations, and in the intuitive handling of the response rate without a transformation. Especially the
latter lowers the hurdle of applying a basket trial because it prevents the impression to non-statisticians of a black box behind the transformations. From our point of view the design by Fujikawa et al. (2020) is the most traceable approach as it uses the properties of the beta-binomial model and shares the information among baskets based on pairwise similarities. The similarity is assessed by the Jenson-Shannon divergence, which we consider as a singular point of critique because it only takes values on $[0.307, 1]$. We rank the similarity or divergence measure as a promising aspect of further research in the optimization of the design by Fujikawa et al. (2020). The hypothesis-driven approaches (Lyu et al., 2020; Zhou & Ji, 2020) facilitate decision-making by directly addressing the hypothesis. Yet the hypotheses hide the interesting part of the trial, the distribution and characteristics of the basket-individual response rates, because these are only present in the formulation of the hypothesis. Although the hypothesis-driven approaches appear different to the other BHMs, there is an interesting connection between them since a Dirichlet process mixture (DPM), which is applied in the RoBoT design, can be formulated as a mixture of infinite BHMs (Zhou & Ji, 2020; Sethuraman, 1994). The sharing of information stands and falls with the assumption that the common characteristic is prognostic for a successful treatment, which is strong and hard to prove because of many possible genetic aberrations and possible interactions with unknown confounders (Kitsios & Kent, 2012; Renfro & Sargent, 2017). Moreover, only relying on the response rate is not necessarily the optimal way to assess similarity of baskets, especially when only few observations are available. Therefore Chu and Yuan (2018b) as well as Zheng and Wason (2019) enable the incorporation of biomarker information to further improve the sharing component. The biomarker must be related to treatment efficacy though. Nevertheless, sharing of information is the key innovation of the basket trial and the presented sharing techniques based on the response rate already proved beneficial characteristics (T1E, power, samples size) compared to independent analyses of the baskets.

The interim futility assessment creates a dynamic framework which displays the expectations of modern trials for personalized treatment. The advantage of interim futility assessments is that they focus scarce resources on promising baskets and prevent patients from futile treatment. A downside is their influence on the power to detect promising baskets because multiple futility interim assessments can lead to a higher number of wrongfully stopped baskets. Though, this disadvantage must be considered together with the advantages and benefits for the trial. The time point of the interim assessment is a critical aspect because it describes the amount of observed information and therefore affects the certainty of the decisions. Asymmetrical recruitment to the baskets can challenge the rules of the interim sequence because the weight of better recruiting baskets is increasing and the sharing of information can shift the distribution of the response rate in baskets with fewer enrolled patients. Moreover, a rule which demands a minimum increase of patients can lead to intermittent trials if one basket struggles to constantly recruit patients. Also, the trial design must clearly predefine if information is shared before the interim futility assessment or if each basket is considered individually. The majority of the trial designs propose the former, whereas Liu et al. (2017), Chen et al. (2016), Li et al. (2019), and Zhou et al. (2019) propose the latter. We suggest to share information just before the interim assessment since the sharing of information is the key idea of a basket trial and this is also what is done in the final analysis. The futility techniques all depend on the so far observed data and they all display the same data in a different way. The decisions they take depend on the technique-specific parameters which can be tuned to come to the same decisions. We regard the posterior predictive probability tool as the most intuitive for interim futility assessments because it directly refers to the final analysis and there is no threshold for the number of observations per basket because this limits the dynamical aspect of the trial design. On the other hand, the posterior probability always expresses the current knowledge; however, with different certainty depending on the number of observations and hence the threshold $\tau_F$ must be adapted over time to achieve similar certainty at each interim assessment. Apart from the already known techniques, we suggested to also consider the conditional power as an interim futility tool in frequentist trial designs. The interim futility assessment is an optional component and should be applied to the best of the trial’s purpose and its goal.

The same holds true for the interim efficacy assessment. This component is supposed to detect efficacious baskets early; however, this requires an overwhelming response because the number of observations is rather low and the threshold to declare early efficacy is high. Nevertheless, it leaves the door open for a positive surprise. We consider the efficacy interim as the component with the least beneficial characteristic for the basket trial, since in the context of exploratory trials, the goal is not to declare early efficacy but to filter baskets with potentially promising behavior. In the confirmatory setting, the interim efficacy assessment can serve as a gateway for accelerated approval in baskets with unmet medical need. We therefore suggest to incorporate the efficacy stop component if it is feasible, while considering the inclusion of this option as more important in a confirmatory setting than in the exploratory setting. In the latter, we argue that the focus should rather lie on the sharing of information and the futility assessments. The rule for the time points of interim efficacy
assessments must be prespecified and we advise to harmonize them with the interim assessment then filters both favorable and unfavorable baskets and helps to keep study logistics as simple as possible with one combined time point that induces changes. The interim futility and efficacy assessment anticipate the final analysis, so they should be consistent with what is planned for the final analysis. Consequently, we propose to keep the reference value (for the response rate) constant throughout the trial which means that interim futility, efficacy, and final analysis use the same reference value. We emphasize this aspect because some designs (cf. Section 3.2) work with a reference response rate and a target response rate following a Simon two-stage design with null and alternative hypothesis. However, we have seen that the reference value is interchangeable with the applied threshold. Therefore, a consistent presentation is not only possible but underlines the well thought choice of tools for each component, and highlights the determination to build a transparent and understandable basket trial.

The final analysis aggregates all observed information and presents the results of the trial. The trial team derives decisions according to the goal (confirmatory, detection of promising baskets, PoC) of the trial, respectively. The final analysis depends on the trial design, especially on the sharing technique. In the pool all or nothing designs, the evaluation is either done individually for each basket or for all baskets together. The latter then evaluates the treatment in general, irrespective of the baskets, and targets a market approval of the treatment for all baskets. In that case, the sharing results in a fusion of baskets. In contrast, the Bayesian designs with intermediate sharing allow a final interpretation for each basket. Interestingly, the PoC final analysis does the same, but aggregates the characteristics of each basket into one probability based on which a proof of the treatment concept in at least one basket is claimed.

From our point of view, the modularity allows to individually tailor a trial consisting of an arbitrary combination of all four components with known and innovative tools for each of them. This will hopefully facilitate the application of basket trials in hands-on medical research as the components are better accessible in a stepwise approach and existing tools can be looked up in this paper. As for other trial designs, one demands to know the characteristics of the basket trial regarding precision and error rates. Therefore the TIE control (marginal for each basket individually and/or family-wise for the complete trial) and the power to achieve the study goal are relevant issues of a basket trial. Both depend on the arrangement of the components, the applied tools and the number of baskets. Each contribute to the number of decisions that are made throughout the trial and therefore to the number of wrong decisions. This problem is generally known in clinical trials as multiple testing problem. Likewise, it is important in basket trials, irrespective of the application of frequentist or Bayesian tools, because decisions are made with both techniques. The false positive decision rates (basket-specific and family-wise) are characteristics to describe the magnitude and the quality of a basket trial design. Basket trials are prone to many sources that could increase the error rates, for example, the interim pruning of baskets can be considered as cherry-picking, and the sharing among baskets might draw actually futile baskets over the finish line, leading to a false claim of efficacy. The control of errors must be considered in the context of the trial purpose. In exploratory designs, the aim is rather to detect the promising baskets and hence it is important to known how likely a futile basket is wrongly declared promising. This metric is given by the basket-specific TIE (cf. Chu & Yuan, 2018a, Fujikawa et al., 2020). On the other hand, in the confirmatory setting, one rather aims to prevent to declare the treatment promising while in truth it does not work in any basket. This is then covered by the family-wise TIE (cf. Chen et al., 2016, Cunanan et al., 2017b). Hence, the control of errors cannot be evaluated globally for all basket trial designs. However, the error rates should be assessed during the planning of a new basket trial. This control is barely possible in an analytical way since analytic calculations are likely to fail with increasing complexity of the trial design (number of baskets, number of interim futility and/or efficacy, time point and technique of sharing). Thus, the underlying type I error rates can often only be approximately estimated via simulations when calibrating the trial design. Also, the TIE control must be kept in balance with the power characteristic, because a strong TIE control might be appealing at first sight, but without sufficient power to detect any activity of the investigated treatment, it is not worth the effort to initiate a basket trial.

To conclude, we state that a good basket trial stands out due to a high (or even optimal) discriminatory power regarding heterogeneous baskets and subgroups of them, and with regard to a high power in the separation of futile and promising baskets. This should therefore be the benchmark and motivation for ongoing and further research. The field of research is wide as it addresses an optimal arrangement of the components, the improvement of existing tools and the development of new methods. Moreover, innovations must prove at least that their performance is similar to existing methods, which reveals a current blind spot in literature. A broad comparison of proposed trial designs with each other is not available. This is lacking because especially the most recent techniques are complicated and were developed in parallel such that they did not reference each other when they were published. However, for the application of the innovative methods and designs in practice, it is essential to know the advantages and disadvantages compared to other options to ultimately
justify the choices made. Another interesting aspect of research is the incorporation of further biomarker information in basket trials. Apart from the BLAST design by Chu and Yuan (2018b), all other presented designs assess the patients and baskets only on the outcome of the primary endpoint. However, it seems reasonable that many factors play a role for the outcome and also for the similarity of baskets. These factors can be displayed by biomarkers. The BLAST design models these as a trajectory over time additional to the responses, and Yin et al. (2020) proposes a biomarker enriched basket design with focus on biomarker-cutoff determination and a subsequent enriched BHM. Trippa and Alexander (2017) propose a design with several (negative and positive) biomarkers and an adaptive randomization. Similarly, does Ventz et al. (2017) focus on a master trial design which is separated into a Bayesian biomarker subpopulation part and a subsequent frequentist identification of cancer types that work better with biomarker-driven treatments compared to standard of care. This approach poses the opportunity for further investigation and specification directly for basket trials. Xu et al. (2019) follow a similar idea with subpopulation finding based on a utility function and adaptive allocation of treatments based on observed responses. However, in accordance with Wathen and Thall (2017), they admit the limited benefit of adaptive allocation. The use of biomarkers is also proposed in Lin et al. (2020), who present a late phase group-sequential design for the analysis of multiple subgroups. The authors do not explicitly assign their design to basket trials. However, they offer a design that can be adapted to a late phase Bayesian basket trial with explicit control of the family-wise error rate under the global null hypothesis for a time-to-event endpoint. They therefore contribute a valuable starting point for the required research especially in late phase basket trials. Progress in this field will fill the so far empty category of confirmatory basket trial designs that apply Bayesian tools. The transfer of the basket trial idea to other phases of the drug development is an appealing field of further research. This also includes phase I dose finding trials with the goal to determine appropriate dose levels for later phases. Also, seamless phase I/II designs with focus on both, toxicity, and efficacy, are of interest as for example the just recently proposed work by Lin et al. (2021). Basket trials for dose finding or for late phases impose an increased complexity compared to the phase II designs. While phase II trials mostly apply binary response variables, the dose-toxicity curves in modern dose finding trials and the time-to-event endpoints with survival and hazard functions in phase III introduce more methodological challenges, especially when information between baskets shall be shared. Future research on basket trial designs and respective methods for dose finding and for late phase trials is needed. The so far available phase II designs can thereby inform the development of novel phase I/II or phase II/III seamless designs.

This work presents trial designs and existing statistical tools for basket trials. The presented methods display the most relevant techniques in the ongoing evolution of basket trials from our perspective. We categorized the trial designs and refined the components that make up a basket trial. We presented the evolution of methodologies in a uniform manner, pointed out differences and connections, and recapitulated the current literature to our best knowledge as of late 2020. Finally, we want to underline that the purpose of this paper is not to remain static, but to be updated and extended in the future as new innovations for basket trials are being published.

ACKNOWLEDGMENTS
The authors thank the anonymous reviewers for their profound comments. These were considerably helpful and an inspiration during the revision of this manuscript.

CONFLICT OF INTEREST
The authors have declared no conflict of interest.

ORCID
Moritz Pohl https://orcid.org/0000-0002-2801-950X
Johannes Krisam https://orcid.org/0000-0003-4092-7874
Meinhard Kieser https://orcid.org/0000-0003-2402-4333

REFERENCES
Asano, J., & Hirakawa, A. (2020). A Bayesian basket trial design accounting for uncertainties of homogeneity and heterogeneity of treatment effect among subpopulations. Pharmaceutical Statistics, 19, 975–1000.
Ashley, E. A. (2016). Towards precision medicine. Nature Reviews Genetics, 17, 507–522.
Bayes, T., & Price (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. Philosophical Transactions of the Royal Society of London, 53, 370–418.
APPENDIX

A.1 Calculation of posterior predictive probability

Posterior predictive probability (PPP) is used as a decision criterion in Fujikawa et al. (2020). $p_i$ is beta distributed and the number of future responses $r_i$ follows a binomial distribution with parameters $n_2$ and $p_i$. The final number of observations is $n = n_1 + n_2$, $n_1$ is the number of observations at the interim futility assessment, and $r^*$ is the number of responses including the shared information. Constant values without further specification are denoted by $c_i$ and $\tilde{c}_i$. The probability of future responses is

$$
\int_0^1 f(r_i|p)f(p|D)dp = 
$$

$$
= \int_0^1 c_i p^{i_1}(1 - p)^{n_2 - r_i} \cdot \frac{1}{B(a + r^*, n_1 - r^* + b)} p^{a + r^* - 1}(1 - p)^{n_1-r^*+b-1}dp = 
$$

$$
= \frac{c_i}{B(a + r^*, n_1 - r^* + b)} \cdot \int_0^1 p^{a + r_i + r^* - 1}(1 - p)^{b + n_1 + n_2 - r^* - r_i - 1}dp = 
$$

$$
= \tilde{c_i} \cdot B(a + r_i + r^*, b + n_1 + n_2 - r^* - r_i)
$$
and consequently the PPP is a sum over Beta functions

\[
\sum_{\{r_i: \text{final analysis successful}\}} \int_0^1 f(r_i | p_i) f(p_i | D) dp_i = \sum_{\{r_i: \text{final analysis successful}\}} \bar{c}_i \cdot B(a + r_i + r^*, b + n_1 + n_2 - r^* - r_i).
\]

### A.2 Workflow of a basket trial

![Workflow diagram](image)

**FIGURE A1** 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

### A.3 Information sharing techniques, connections and categories
FIGURE A2  Each box with name, author, and year of publication represents a sharing technique. 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.
FIGURE A3  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