Efficiency of a Randomized Confirmatory Basket Trial Design Constrained to Control the Family Wise Error Rate by Indication

Linchen He[1,2], Yuru Ren[1], Han Chen[1], Daphne Guinn[3,4], Deepak Parashar[5,6], Cong Chen[7], Shuai Sammy Yuan[7,8], Valeriy Korostyshevskiy[1], Robert A. Beckman[1,9]

[1]Department of Biostatistics, Bioinformatics and Biomathematics, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC USA

[2]Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY, USA

[3]Program for Regulatory Science and Medicine, Georgetown University, Washington, District of Columbia, USA

[4]Department of Pharmacology and Physiology, Georgetown University, Washington, District of Columbia, USA

[5]Statistics and Epidemiology Unit & Cancer Research Centre, Warwick Medical School, University of Warwick, Coventry UK

[6]The Alan Turing Institute for Data Science and Artificial Intelligence, The British Library, London UK

[7]Biostatistics and Research Decision Sciences, Merck & Co., Inc., Kenilworth, NJ USA

[8]Kite Pharma, a Gilead Company, 2400 Broadway. Santa Monica, CA 90404

[9]Department of Oncology, Lombardi Comprehensive Cancer Center and Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC USA

Corresponding author:

Linchen He, Department of Biostatistics, Bioinformatics and Biomathematics, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC USA

Email: lh1790@nyu.edu

Abstract

Basket trials pool histologic indications sharing molecular pathophysiology, improving development efficiency. Currently basket trials have been confirmatory only for exceptional therapies. Our previous randomized basket design may be generally suitable in the resource-intensive confirmatory phase, maintains high power even with modest effect sizes, and provides nearly k-fold increased efficiency for k indications, but controls false positives for the pooled result only. Since family-wise error rate by indications (FWER) may sometimes be required, we now simulate a variant of this basket design controlling FWER at $0.025k$, the total FWER of k separate randomized trials. We simulated this modified design under numerous scenarios varying design parameters. Only designs controlling FWER and minimizing estimation bias were allowable. Optimal performance results when $k = 3,4$. We report efficiency (expected # true positives/expected sample size) relative to k parallel studies, at 90% power ("uncorrected") or at the power achieved in the basket trial ("corrected", because conventional designs could also increase efficiency by sacrificing power). Efficiency and power (percentage active indications identified) improve with higher percentage of initial indications active. Up to 92% uncorrected and 38% corrected efficiency improvement is possible.

Even under FWER control, randomized confirmatory basket trials substantially improve development efficiency. Initial indication selection is critical.

Keywords: *confirmatory basket trial, adaptive design, family-wise error rate, power by indication, cost-effectiveness*

INTRODUCTION

Molecular oncology has led to increasingly numerous biomarker-defined niche indications.[1] For example, lung cancer now includes several small subgroups with distinct therapies. This may lead to clinical trial enrollment challenges, and to additional development expense and delay. Conversely, a targeted therapy may have potential application in numerous indications, as well as in multiple combination settings, creating a large number of potential clinical hypotheses for testing with finite resources.

We previously discussed the need for increased development efficiency in design of proof of concept studies and associated Go-No Go decisions, given the large number of potential hypotheses worthy of testing, which may strain available resources.[2] Efficiency is a measure of the utility of a study per unit of resource (number of trial participants and/or financial cost) expended in the study or *as a consequence of the study*. For example, in the instance where a Phase 2 proof of concept study gives a false positive result, Phase 3 resources will likely be expended as a consequence of the proof of concept study in a futile effort to confirm the false positive result, and this outcome can be used to penalize the

efficiency value of the proof of concept study design, probability weighted by its Type I error rate).[2] The efficiency of a proof of concept study design has therefore been expressed as the probability of finding a true positive proof of concept (utility) divided by two cost terms: the number of trial participants utilized in the proof of concept study itself and the probability weighted number of Phase 3 trial participants for Phase 3 trials resulting from true and false positive results. To calculate this quantity, one needs an estimated probability distribution that a therapy entered into the proof of concept trial from a population of therapies will be active in the state of nature.[2]

Efficiency can be defined differently in different contexts in a fit for purpose fashion. It is a fundamental metric for judging the value of clinical trial designs. In contrast to power, it incorporates cost (measured financially or in trial participants). It is always possible to increase the study power, by simply increasing its sample size, but the cost of doing so continually rises with increasing power, and the point where one reaches diminishing returns in seeking power is subjective, although governed by traditions. By incorporating a probabilistic estimate that the therapy is truly active, as well as setting a defined Type I error threshold and a defined target, efficiency metrics blend such concepts as positive and negative predictive

value into a summary of the return (in useful drug development knowledge) on investment.

Importantly efficiency can be calculated for a single study or for an ensemble of studies, or even across a clinical development portfolio. In the work on proof of concept studies, the realistic case was considered in which budgetary constraints were applied to a portfolio of proof of concept opportunities of equal merit. The budget (in available dollars or trial participants) was insufficient to perform traditional proof of concept studies testing all of the proof of concept hypotheses. In this setting, a surprising result was found: efficiency was higher if the proof of concept studies were reduced in size, cost, and power compared to the traditional 80% power, allowing a larger number of hypotheses to be tested within the fixed budget. Otherwise, there was too high an opportunity cost of not testing credible hypotheses under the budgetary constraint. There was the possibility of not testing hypotheses that might have resulted in true positives, an event which we termed a Type III error.[1]

This paper is primarily concerned with the efficiency advantages of basket trial designs in the resource intensive confirmatory phase of development. In all phases of development, less efficient approaches may contribute to the very high

cost of therapy, to long development times delaying availability of therapy, as well as to decisions not to develop drugs for niche indications. Given that the confirmatory phase is the most resource-intensive, efficiency improvements in this phase may have the greatest practical impact.

Master protocols[3,4] can potentially increase the efficiency of drug development, and facilitate development of niche indications. Master protocols include platform trials, umbrella trials, and basket trials. In platform trials, different therapies are perpetually cycled through an ongoing trial, resulting in notable operational efficiencies due to the existence of a common infrastructure, the “platform”. In umbrella trials, multiple therapies are matched to multiple biomarker subgroups within a single traditional organ-system-based indication. Enhanced efficiency comes primarily from the ability to share a common standard of care control over multiple experimental arms. Adaptive randomization, in which randomization probabilities are adjusted frequently based on interim results, may also improve efficiency. These efficiencies are not primarily operational, but rather are inherent in the design.

In a basket trial, traditional indications are grouped together in a basket based on a shared molecular or pathophysiologic characteristic thought to predict utility of a

therapy. These indications may borrow information from each other, or may be frankly pooled, leading to large improvements in development efficiency. This development efficiency is again inherent in the design, arising directly from the fact that multiple indications may contribute to the total sample size if they may be fully pooled. In principle, if k indications are fully pooled, they may all be tested in the pool for the same sample size N that would have been used for the testing of only one indication in a traditional design. Under ideal conditions, a k-fold increase in development efficiency may thus be achieved, a far greater efficiency increment than is available from platform trials or umbrella trials. The challenge in optimizing basket trials to approach this substantial benefit comes from the risk of heterogeneity between indications despite shared biomarkers sometimes conferring similar benefits. Active indications (defined as indications in which the therapy is active) may carry inactive indications along with them to create a positive pooled result. Conversely, inactive indications may dilute the effectiveness of active indications, leading to a negative pooled result. These effects of heterogeneity must be accounted for, and thus the efficiency gain of basket trial designs will be less than k-fold, compared to the ideal case.

Several efficient basket trial approaches use response rate data in the exploratory setting. Most of these are based on Bayesian hierarchical models[5-7] with the exception of one that considers the likelihood that the indications all come from one statistical distribution, versus the likelihood they are best modeled individually.[8]

The first oncology basket trial in a regulatory approval setting was for imatinib, which had already demonstrated extraordinary value in chronic myelogenous leukemia, and had been rationally designed based on considerable scientific evidence.[9] The study was non-randomized and based on response rate, with very small sample sizes. Forty indications were evaluated in less than 200 patients. One approval resulted from 1 response in 5 patients. Note this design did not utilize pooling. In this case, operational efficiencies were realized by consolidating what would have been 40 tiny studies. However, in some settings basket designs have operational challenges. For example, a basket design may require operational cooperation between GI oncology, lung oncology, and breast oncology, three divisions which may lack experience working together or uniform standard operating procedures for tissue collection[4]. Related designs resulted in approvals for transformational drugs designed for alterations in b-raf and ntrk

oncogenes.[10,11] Recently, the immune checkpoint inhibitor pembrolizumab was approved in multiple solid tumors, based on a basket trial pooling patients from these indications with a DNA repair defect resulting in microsatellite instability, utilizing response rate as a primary endpoint.[12] In all of these cases, the drugs and biomarkers were supported by unusually strong scientific evidence, had previously achieved transformational results, and were being studied in underserved indications, and thus were able to merit approval based on single-arm response rate data in relatively small populations.

Motivation for the Research

In contrast, many effective oncology drugs have required rigorous randomized designs in the confirmatory setting. We developed a confirmatory basket trial design[13, 14] which, in addition to being applicable in a single-arm fashion or with response rate endpoints, can also be utilized in a randomized controlled setting using time to event (TTE) endpoints such as progression free survival (PFS) and overall survival (OS). Randomization is generally important in approval of agents whose clinical benefit must be measured based on TTE endpoints, unless the effect size is transformative. We have been interested in developing a basket trial design potentially suitable for the majority of confirmatory settings. We are

particularly interested in the confirmatory phase of development. Because the confirmatory phase is the most resource and time-intensive phase, the marked improvements in efficiency that are potentially available from basket trials can have the greatest impact in efficiently bringing beneficial therapies to patients.

We have previously published a randomized confirmatory basket trial design.[13, 14] Although efficiency was not formally evaluated, an example application was given in which a confirmatory study evaluated 6 indications for a sample size which would be typical for one or at most two typical Phase 3 studies, i.e. a 3-fold or 6-fold increase in efficiency. This original adaptive design resembles a funnel (Figure 1A, see Methods for more details). Indications are carefully selected, and then filtered (removed) in several “pruning steps”, first with data external to the study (i.e. maturing Phase 2 data from the same drug, data from other agents in the class), and then with data from an interim analysis, which may be based on a surrogate endpoint considered predictive of the definitive endpoint (i.e. PFS predicting OS) or on early analysis of the definitive endpoint. The interim analysis is performed on each indication individually to facilitate the pruning of inactive indications. After pruning, a sample size adjustment may be applied to the remaining indications, which are then pooled. The study concludes with a pooled

analysis of the remaining indications based on the definitive endpoint, and is positive if statistically significant benefit is shown. Descriptive statistics including hazard ratio, confidence intervals, and safety data are presented by indication for informal benefit/risk analysis by health authorities. Individual indications may be removed from the pool if their results are inconsistent with the pooled result, at the discretion of health authorities. Ideally, the extent of and defining criteria for this individual indication analysis should be negotiated with health authorities in advance.

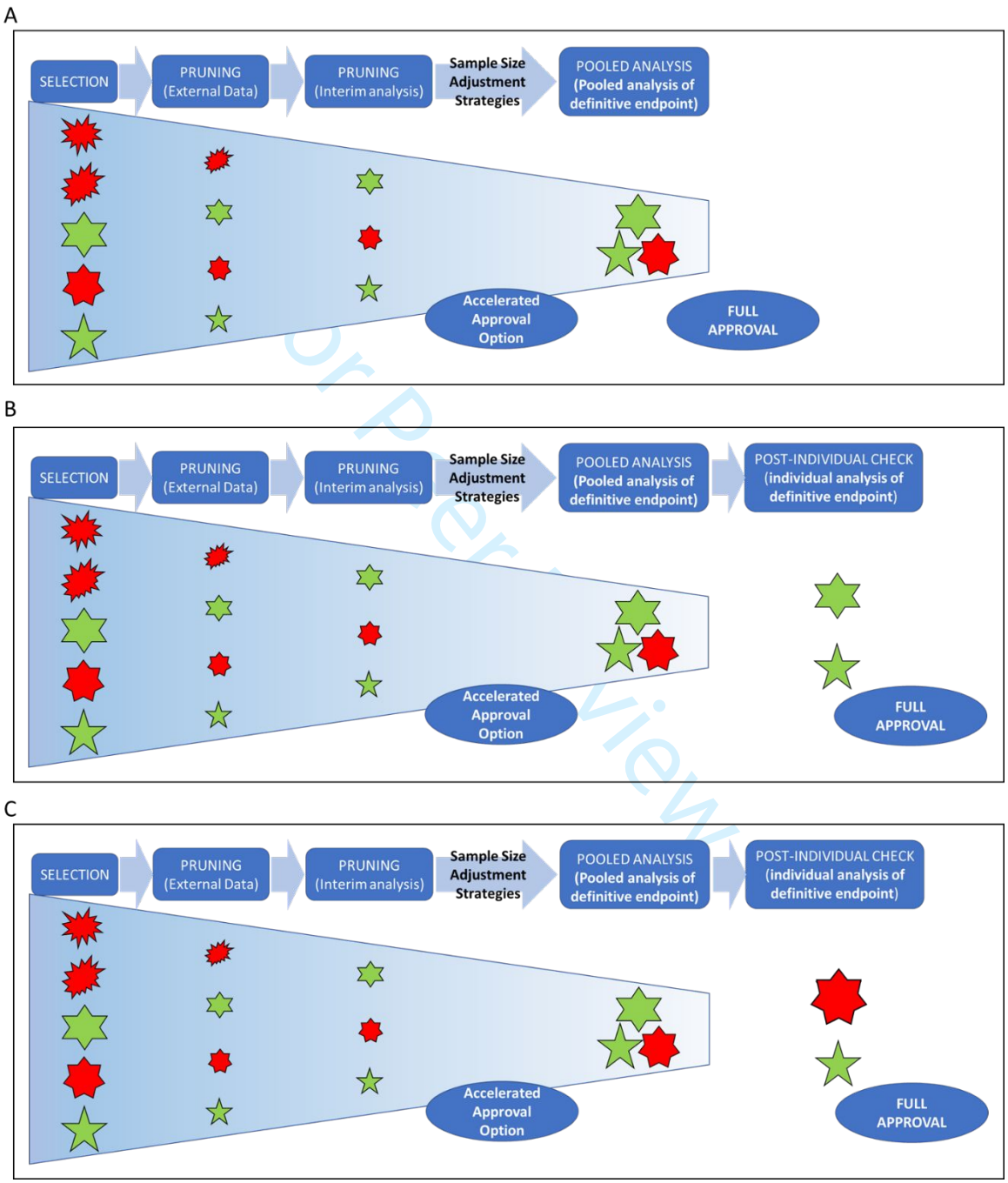


Figure 1. Confirmatory basket trial design (A) with pruning and pooling as in previous studies[13,14], and (B-C) with pruning, pooling, and post-individual check simulated in this study. The designs (1) conduct a basket trial that consists of k tumor indications; (2) prune (remove) indications based on external data; (3) conduct an interim analysis independently for each tumor indication. Indications that meet these interim criteria may in some cases be eligible for accelerated approval, indications that do not are pruned; (4) adjust sample size of remaining indications as needed; (5) conduct a pooled analysis of the remaining indications; (6) in the current design (B-C) only, conduct a prospectively defined post-individual check for each indication involved in the pooled analysis. In previous studies (A), indications that passed the pooled analysis may potentially be eligible for full approval, whereas in the new design utilized in this study (B-C), indications must also pass a simultaneous post-individual check to be potentially eligible for full approval. (B) shows the final checking step successfully removing an inactive indication from the pool as intended. (C) shows a case where the final checking step fails to remove an inactive indication, and instead mistakenly removes an active indication. Active indications are in green, inactive indications in red.

This original design demonstrated a dramatic increase in development efficiency, which we define in this work as (expected number of true positives/expected sample size subject to control of the false positive rate), and maintained acceptable power over a variety of scenarios even with inactive indications in the basket.[13, 14] However, it was designed to control the false positive rate only in the

pool as a whole (Figure 2A, column 3), i.e. a pool which contains one or more truly active indications is considered a true positive. The design does not control the family wise error rate (FWER) by traditional indication subgroups (Figure 2A, column 4; Methods, Supplemental Methods). In this paper, we define active and inactive indications as indications in which the test drug does or does not provide clinical benefit, in the unknown state of nature. FWER by indication subgroup, simply called FWER in this paper, is defined as the probability that one or more inactive indications will be approved by the design. To further illustrate these principles, we show analogous definitions for negative and positive predictive value (NPV and PPV) in Figure 3 (see also Table 1 for definitions). When the indications are carefully selected, as recommended for these designs, the majority should be true positives. Under these conditions, we found that due to the low prevalence of true negatives, NPV was low, and PPV was higher than one minus the FWER (not shown).

As we elaborate in the discussion, formal control of the family-wise error rate by subgroups is not normally required in a confirmatory study, although informal ad hoc subgroup analyses may be performed in the approval setting. Nonetheless, for basket trials, in which the subgroups correspond to traditional indications,

published opinions of authors with past or present affiliation with the European Medicines Agency and other European health authorities[15], as well as our own informal interactions with health authorities, indicate that control of the FWER by indication may at times be recommended for basket trials in confirmatory settings. While this is a subject of ongoing discussion, at the present time there is a practical need for a variation of our original design which provides control of the indication-specific FWER, in case it is required for approval in a particular setting. In the current environment, discussion with relevant health authorities is recommended before deciding whether to utilize the original randomized confirmatory basket design[13, 14] or the new variant to be presented in this paper.

The aims of this present paper are to introduce a variant of the original randomized confirmatory basket design,[13, 14] demonstrate that it controls indication-specific FWER, and characterize its performance in terms of efficiency gains relative to traditional designs ("relative efficiency"), power, FWER, estimation bias, and confidence interval estimation. We observe some power losses with the new design. We note that the efficiency of a traditional design may be improved simply by reducing the power, due to the increasing cost of incremental gains in power as power increases. Hence, if our new design results in some power loss,

the appropriate efficiency comparator is a traditional design with the same reduced power ("corrected relative efficiency").

Cunanan et al documented that a well-known exploratory basket trial design did not control FWER, and, without opining whether this was acceptable in the exploratory setting, advocated for disclosure of performance properties for complex designs.[16] We agree, and characterized our original randomized confirmatory basket trial in this respect, finding that it did not control FWER with acceptable power levels (data not shown). It may still be suitable in those confirmatory settings where control of FWER is not recommended.

In order to control FWER, and also maintain acceptable power levels, we implemented a modification of the initial design in which, whenever the pooled result is positive, each indication is re-checked at low to medium stringency for statistical significance before final approval (Figure 1B and C, Methods). The series of tests, each at lower stringency (higher alpha) is sufficient to control FWER to a prespecified level more stringent than each of the individual checks. As has been shown in biological systems where high fidelity with minimum energy expenditure is required[17,18] repeated testing at lower stringency is more efficient than a single highly stringent test.

We describe the performance of this modified design in extensive simulations, focusing on a scenario with the same TTE endpoint at interim and final analyses. By varying design parameters (Table 1, Methods, Supplemental Methods), we create numerous design variants. We select the variant that maximizes performance, generally judged by relative efficiency. Acceptable design parameters must control FWER to the same level as a system of individual randomized studies for each indication in parallel, i.e. approximately 0.025 multiplied by the number of indications. Further, acceptable design parameters must not introduce bias of greater than 10% in the effect size estimate, and the estimated 95% confidence interval of the effect size must cover at least 90% of simulation runs. For the input parameters to be considered acceptable, these constraints must be met regardless of the number of inactive indications within the basket.

For design parameters meeting these constraints we characterize development efficiency and power as estimated by simulation. We define development efficiency as the expected number of true positive indications identified divided by the expected sample size. “Uncorrected” relative development efficiency is the efficiency of the basket design divided by that of a group of parallel traditional

confirmatory studies, powered at 90%, investigating the same indications. As we do not achieve 90% power by indication in the basket trial, we present a "corrected" relative development efficiency adjusting for the power losses as described above. As enhanced development efficiency can be achieved in conventional designs by reducing power alone,[2] corrected relative development efficiency compares the efficiency of the basket design to parallel conventional studies at the same power.

Power is evaluated by indication (Figure 2C), a more stringent criterion than traditional confirmatory studies, where subgroups are usually not formally powered. Power is therefore the fraction of active indications expected to be qualified for approval by the basket trial. For comparison, we also present the traditional power of the pooled analysis.

We present in Results selected designs that optimize corrected relative efficiency, either subject to a minimum power constraint (recommended development scenarios) or irrespective of power. These designs are characterized as a function of the number of total indications, the number of active indications, and the degree of activity (hazard ratio 0.5 – 0.8).

We discuss overall utility of the design and key learnings for performance optimization. We propose criteria for when control of FWER should and should not be recommended by health authorities in confirmatory studies. Finally, we outline future research topics aimed at addressing other potential concerns with randomized confirmatory basket trial designs.

Methods

Study Design Overview and Design Parameters

Consider a randomized confirmatory basket trial of an experimental therapy that consists of k tumor indications. For each indication, we perform 1:1 randomization (experimental vs. control), with a TTE variable as the primary endpoint of interest and n events per indication.

Figure 1B presents the current study design. We assume an interim analysis is conducted on each tumor indication individually at a common information time $t \in (0,1)$ based on nt events for all tumor indications, which we assume is also the same actual time. The study designer should choose sample sizes to make this approximately true, but may need to conduct several interim analyses and sample size adjustments in practice. We assume a common interim bar α_t (one-

sided nominal Type I error rate) across tumor indications for simplicity. A tumor indication is "pruned" from the study at interim analysis if it does not meet the bar for pooling. Remaining indications proceed to the pooled analysis. After the interim analysis, we adjust the sample size for the remaining tumor indications to account for lost sample size due to pruning. We consider three sample size adjustment designs[13, 14]:

Table 1. Glossary of Terms

Design parameters	Descriptions
The state of nature	
g	The number of active indications at the beginning
θ_i	The hazard ratio of experimental arm vs. control arm for the definitive endpoint for indication i
Study design input parameters	
k	The number of tumor indications at the beginning
D	Sample size adjustment strategies
α_t	A common bar for the interim analysis to prune inactive indications
α	False positive rate for the pooled analysis
β	False negative rate for the pooled analysis
α_{post}	A common bar for the post-individual tests
Outcome measurements	
m	The number of indications in the pooled analysis. $m \leq k$
α_{net}	The probability of the basket trial passing the pooled test and at least one false positive indication passing the post-individual test for a given value of g
Family-wise error rate (FWER)	The maximum probability of the basket trial passing the pooled test and at least one false positive indication passing the post-individual test for any g. The possible values of g indicate the "family" for which FWER is defined.

Power by indication	The probability of an active indication passing the interim test, the pooled test, and the post individual test
Power by basket	The probability of a true positive basket (with one or more active indications) passing the pooled analysis
Efficiency	The ratio of average number of active indications that pass the post-individual tests divided by the average sample size
Uncorrected relative efficiency	The efficiency of the basket design divided by that of a group of parallel traditional confirmatory studies, powered at 90%, investigating the same indications.
Corrected relative efficiency	The efficiency of the basket design divided by that of a group of parallel conventional studies at the same power by indication as observed for the basket design, investigating the same indications.
Negative predictive value (NPV) by indications	The proportion of true negative indications among all negative indications not passing the interim test, the pooled test, or the post individual test
Negative predictive value (NPV) by basket	The proportion of true negative baskets among all negative baskets without any indication passing the interim test, the pooled test, or the post individual test
Positive predictive value (PPV) by indications	The proportion of true positive indications among all positive indications passing the interim test, the pooled test, and the post individual test
Positive predictive value (PPV) by basket	The proportion of true positive baskets among all positive baskets with at least one indication passing the interim test, the pooled test, and the post individual test

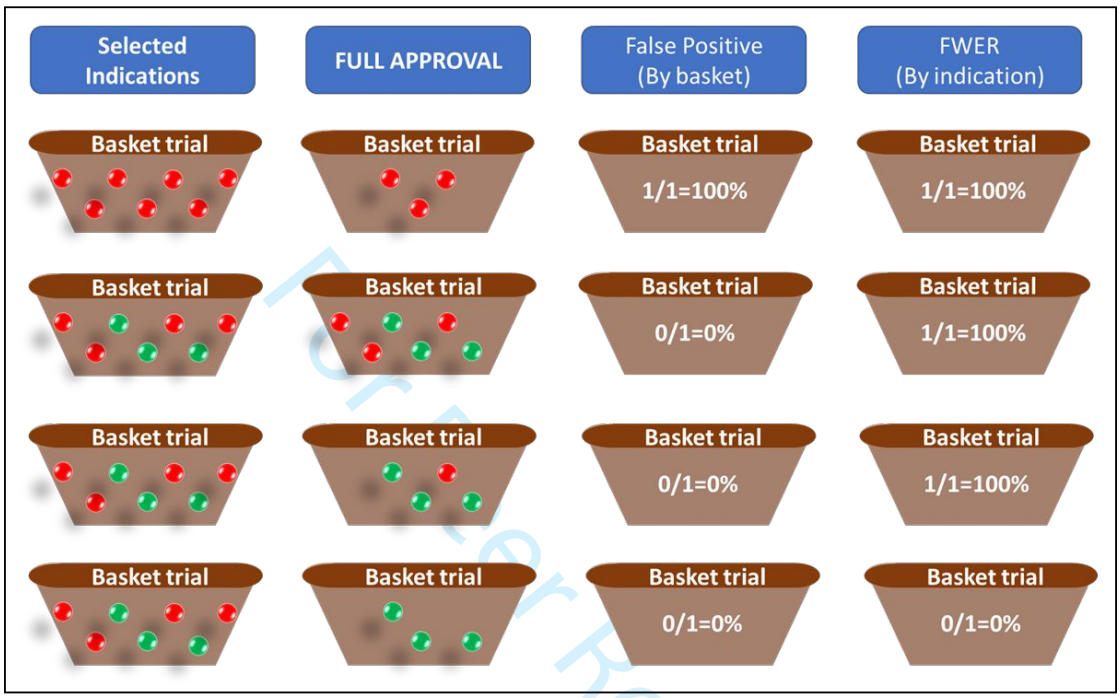
1. Design one (D1): No sample size adjustment.

2. Design two (D2): Aggressive sample size adjustment to replace all originally planned events in the pruned indications.

3. Design three (D3): Moderate sample size adjustment to replace future originally planned events after the interim analysis in the pruned indications.

Although endpoints for pruning and pooling may be different, in this work we consider the same endpoints only. Denote (α,β) as the false positive and negative rates for one-sided hypothesis testing in the pooled population. The adjusted false positive rate α^* to control the false positive rate for the pooled analysis at the desired level will be calculated (see Supplemental Methods) for each strategy with respect to the global null hypothesis that all indications are inactive (Figure 2A, Supplemental Methods).[13,14] Suppose m tumor indications are included in the pooled analysis ($m \geqslant 1$). When a basket passes the pooled analysis, a prospectively defined individual post-pool analysis, examining each of the m tumor indications remaining in the pool after all pruning is complete, is conducted, termed a "post-individual test". We also assume a common prospective post-individual bar α_{post}, varied independently of α_t. Indications that survive the post-individual test may be eligible for full approval.

A



B

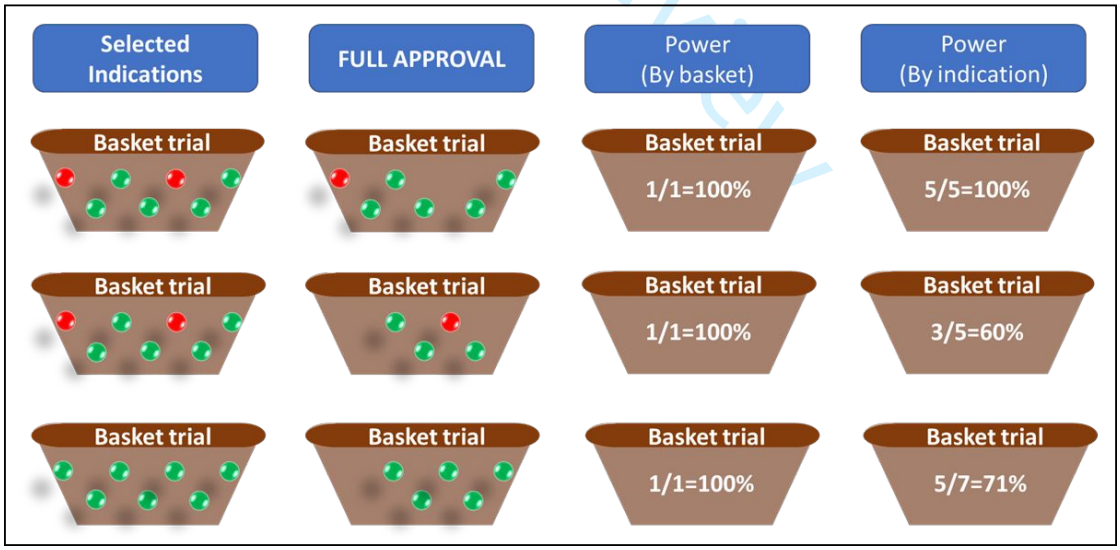
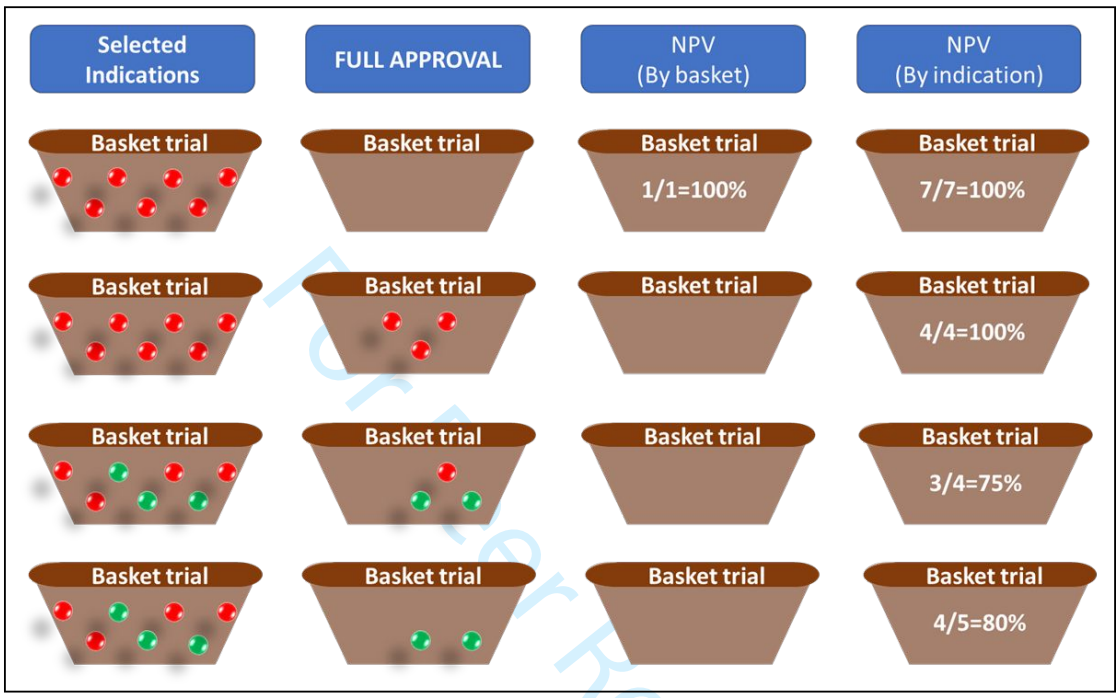


Figure 2. Measurements of false positive rate (type I error) (A); and power (B) showing the difference between criteria used in this study and less stringent criteria. Examples are shown for illustration. Each row represents an example and marbles represent the active (green) and inactive (red) indications. The left-hand column represents the initial selected indications. The second column represents the approved indications. The third column represents traditional criteria (by basket), and the fourth column represents the more stringent criteria (by indication) used in this study. (A) This study is designed to control the family-wise error rate (FWER) by indication subgroup (fourth column) rather than the false positive rate in the pool against a null hypothesis where all indications are inactive. In the third column, a false positive is scored only when a basket is approved containing only inactive indications (false positive rate by basket, as in the original design[13, 14]. Thus, the numerator for FWER is the number of approved baskets with zero active indications and the denominator is the number of approved baskets. In the fourth column, a false positive is declared if a single inactive indication is approved. Thus the numerator for FWER is the number of approved baskets with one or more inactive indications and the denominator is the number of approved baskets. Note the term "basket" means the basket which contains a collection of marbles (indication). (B) Power results are evaluated by indication in this study rather than by basket. Considering the power by basket (third column), the approved basket can be considered as a true positive if it has at least one active indication. Thus numerator is the number of "selected indications" baskets containing one or more active indications that result in an approval, and the denominator is the number of selected indications baskets containing one or more active indications. In the fourth column, the power by indication is defined as the percentage of active indications approved, and can be calculated as 100%, 60%, and 71% for the three examples, respectively.

A



B

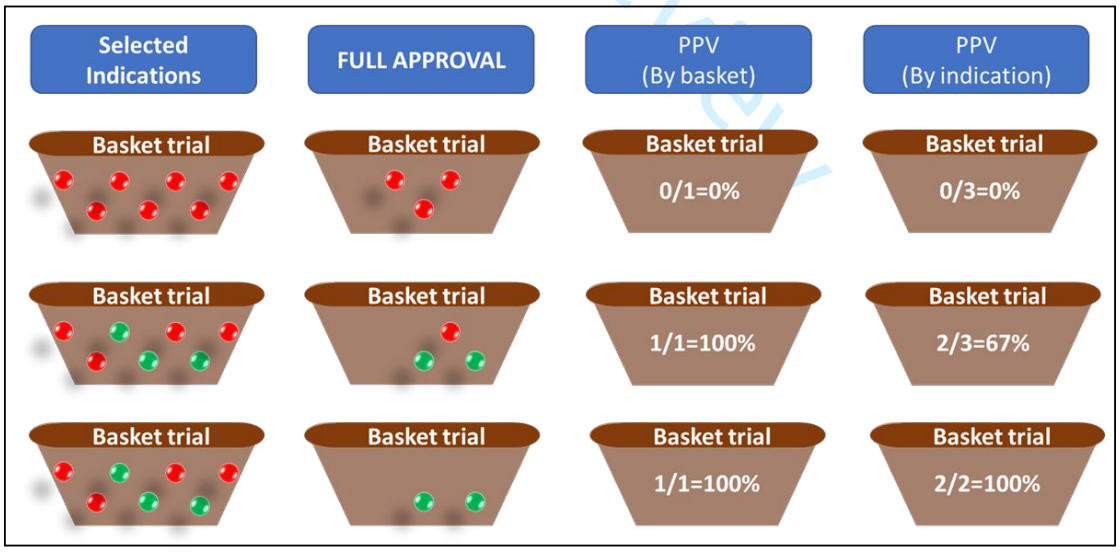


Figure 3. Measurements of negative predictive value (NPV) (A); and positive predictive value (PPV) (B) showing the difference between criteria by indication and by basket. See definitions of NPV and PPV in Table 1. Examples are shown for illustration. Each row represents an example and marbles represent the active (green) and inactive (red) indications. Baskets denote the collection(s) of indications within the container(s) shown. The left-hand column represents the initial selected indications. The second column represents the approved indications. The third column represents the NPV/PPV (by basket), and the fourth column represents the NPV/PPV (by indication) used in this study. (A) In the third column, a NPV is scored only when a basket is not approved, and a basket is considered negative only if all indications are negative. Thus the numerator is the number of baskets not approved that have all indications negative and the denominator is the number of baskets not approved. In the fourth column, a NPV is calculated based on the indications that are not approved and whether they are active or inactive. For baskets that are not approved, the numerator is the number of inactive indications in the basket, and the denominator is the number of indications in the basket. (B) PPV results are evaluated by indication in this study rather than by basket. In the third column, a PPV by basket is scored only if a basket is approved. A basket is considered positive if it has at least one active indication. Thus the numerator is the number of approved baskets that have at least one active indications and the denominator is the number of approved baskets. A PPV by indication is defined if one or more indications are approved. In the fourth column, the PPV by indication is defined as the percentage of active indications approved among all approved indications, and can be calculated as 0%, 67%, and 100% for the three examples, respectively.

Type I error evaluation

In this basket trial setting (Figure 1B) designed to examine multiple indications (k in number), there is no clear analog of conventional Type I error rates if we consider Type I error by indication, since the possibility of committing a Type I error may occur for tests of each indication. We consider the familywise error rate (FWER) by indication, which is defined as the probability of at least one false positive indication getting approved irrespective of the number of active and inactive indications, defined respectively as indications in which the drug provides or does not provide clinical benefit in the unknown state of nature. This FWER considers a family of null hypotheses in which one or more of the k indications are inactive, i.e. 2^k-1 null hypotheses. The FWER is progressively controlled by three successive pruning steps, none of which individually provide Type I error control as stringently as all three in sequence. Considering a basket trial that consists of k tumor indications, these three analyses must be passed for an indication to be approved:

1. Interim analysis: For each of k tumor indications independently, an interim analysis prunes (removes) inactive indications. We assume a common bar

α_t (in terms of the one-sided nominal Type I error rate) for all tumor indications for simplicity.

2. Pooled analysis: For all remaining indications that pass the interim analysis, a pooled analysis is performed relative to the null hypothesis that all indications in the pool are inactive. The adjusted nominal level α^* is used (further details in Supplemental Methods).[13,14]
3. Post-individual check: For each indication that passes the pooled analysis, a prospectively defined post-individual check determines whether an indication may be eligible for full approval. We assume a common bar α_{post}, which is varied independently of α_t.

The three analyses are conducted at the nominal levels α_t, α^*, and α_{post}, however, none of these nominal levels quantifies the Type I error of the entire trial. Rather than control the false positive rate of any of three tests at the level of 0.025 one-sided per indication, we evaluate the FWER of the entire trial according to the cumulative effect of three sequential tests. To evaluate the FWER, we design comprehensive simulation studies, where a range of nominal levels α_t and α_{post} is selected (Supplemental Table S1) and the adjusted nominal level for testing the pool α^* is calculated for each strategy. For a given set of design parameters and

hazard ratio of active indications, the overall Type I error resulting from the cumulative effect of the 3 steps for a given value of g is $\alpha_{net}(g)$, where g is the number of active indications ($0 \leqslant g \leqslant k$). We then vary g from 0 to k with all other parameters constant to get the corresponding value of FWER:

$$\text{FWER} = \max\{\alpha_{net}(g): g = 0, 1, \ldots k\}$$

FWER is considered controlled to the level of α_{target} if $\alpha_{net} \leq \alpha_{target}$ for a given set of input parameters and all g from 0 to k, specifically, $\alpha_{target} = 0.025k$ in this study.

Outcome Measurements

We consider the power by indication, which evaluates the proportion of active indications passing the post-individual tests. We also examine the efficiency, defined as the ratio of average number of true positive indications passing the post-individual tests divided by the average sample size, when subject to control of the FWER by indication. Other measurements include the coverage of the confidence interval (CI) of the hazard ratio (HR) and bias of the estimated HR. In evaluating outcome measurements, we threshold the FWER at the level $0.025k$, the coverage of the 95% CI of the HR as greater than 90% of the simulation runs, and the bias of estimated HR less than 10%. Design parameter combinations

which cannot meet these criteria irrespective of the value of g in our simulation are not allowed. Finally, we examine power by indication and efficiency for each allowed design and utilize a ratio to compare efficiency relative to a reference design.

The reference design for the uncorrected relative efficiency calculation assumes parallel, independent Phase 3 designs planned for each indication with the false positive and false negative rates $(\alpha_{ref},\beta_{ref})$ set to $(0.025,0.1)$.

For calculation of relative efficiency corrected for power losses, we first determine the power by indication of a basket study with the same design parameters and inputs, then set β_{ref} equal to 1 minus this power, while maintaining α_{ref} at 0.025, and finally proceed as for the calculation of relative efficiency above. This correction is necessary because the efficiency of the reference designs are higher when run at lower power[2]. Therefore, comparison to a reference design at the same power is a more stringent, balanced relative efficiency comparison.

Other measurements include the coverage of the confidence interval for HR and estimation bias for HR. Specifically, the CI coverage for HR is defined as the probability that the estimated 95% CI for the HR covers the true HR given that the

individual test passed for that indication. Representing the relative difference in the true HR and the estimated HR, the estimation bias of HR is defined as the ratio of estimated average HR and true pooled HR for those indications that pass the individual tests minus 1.

Simulation study

In our simulation study, we used parameter values that are summarized in supplemental Table S1. We assume that for each indication the true hazard ratio (HR) $\theta_i; i = 1,...,k$ is either at a null value $\theta_0 = 1$ or at an active value $\theta_a \in \{0.5, 0.6, 0.7, 0.8\}$ and the true number of active indications is $g = 0,1,...,k$. For simplicity, we consider an exponential model for the distribution of event times and do not consider censoring. We fix $\alpha = 0.025$ and vary design parameters k $(3,4,5,6)$ and β $(0.025,0.05,0.1,0.2)$. Consequently, the total sample size in the pooled population (kn) can be calculated as $kn = 4(Z_{1-\alpha} + Z_{1-\beta})^2/[(log\theta)^2]$. For simplicity, we set a common information time for the interim analysis $t = 0.5$, so that each indication should accrue nt patients for interim analysis. We note that the in practice the indications should be chosen so that they are projected to reach nt events at approximately the same time to avoid the operational inconvenience of

multiple interim analyses and sample size adjustments. Denote n_{i1} ($n_{i1} = nt$; $i = 1,...,k$) as the sample size for i-th indication at the interim analysis. The sample size for each indication at pooled analysis, denoted as n_{i2} ($i = 1,...,m$), should accrue as follows:

1. $n_{i2} = n$ under D1;
2. $n_{i2} = \frac{kn}{m}$ under D2, which is greater than the sample size under D1;
3. $n_{i2} = n\left(t + \frac{k(1-t)}{m}\right)$ under D3, which is greater than the sample size under D1 but smaller than the sample size under D2.

With these specifications, the total number of patients enrolled in the study can be calculated as $(k-m)n_{i1} + mn_{i2}$ if each indication has the same planned number of patients. We further explore the design for values of α_t and α_{post} of ($0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4$), setting these two parameters independently. For each setting, we generate 10000 simulated trials for the evaluation and comparison.

Simulations and analysis are performed using R (version 3.6). Source code and details of simulation settings and outcome measurements are available in Supplemental R codes.

RESULTS

We explore power and efficiency with various input parameters (Table 1, Supplemental Table S1). We control FWER at FWER $\leqslant 0.025$ for all values of g, the number of active indications, from 0 to k inclusive. FWER is more difficult to control with increasing number of tumor indications and increased therapeutic effect size, reflecting greater heterogeneity (examples shown in Figure 4; more details included in Supplemental Table S2). FWER α_{net} is generally maximal when g is approximately half of k, again reflecting maximal heterogeneity.

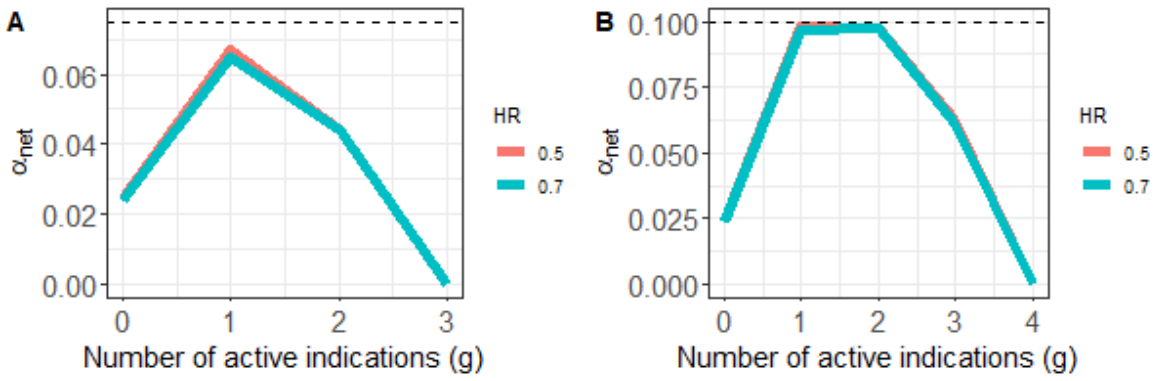


Figure 4. Controlled FWER at FWER $\leqslant$ 0. 025k for all values of g ($g = 0,...,k$). (A) α_{net} is shown on the y axis, and the number of active indications on the x axis. α_{net} = FWER is controlled at $\leq 0.025k$ (dotted horizontal line) for k =3, $HR = 0.5$ (orange; Design 3, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.35$, $\alpha_{post} = 0.05$, $\alpha^* = 0.0089$) and $HR = 0.7$ (blue; Design 3, $kn = 409$, $\beta = 0.05$, $\alpha_t = 0.4$, $\alpha_{post} = 0.05$, $\alpha^* = 0.009$). (B) same as (A), for $k = 4$, $HR = 0.5$ (orange; Design 2, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0094$) and $HR = 0.7$ (blue; Design 3, $kn = 483$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0075$).

We provide recommended optimal design parameters and associated performance results (Figure 5, Supplemental Figure S1). These optimal design parameters determined by simulation depend on the scenario studied and thus vary from panel to panel in the Figures. The legends list these parameters. These recommendations consider both corrected relative efficiency and power. Although corrected relative efficiency is arguably

the most fundamental metric, many sponsors will prefer to have power within an "acceptable" range for a confirmatory study. We also provide design parameters that optimize corrected relative efficiency irrespective of power (Figure 6, Supplemental Figure S2) and show associated performance results. All simulation results are available in Supplemental Table S2.

Decreased power and/or efficiency if the majority of indications are inactive

Power and/or efficiency increases with the proportion of indications that are active. (Figures 5-6, Supplemental Figures S1-S2). Performance deteriorates when multiple inactive indications are present in the basket.

As the hazard ratio of positive indications increases, indicating a decreased therapeutic effect size, the corrected relative efficiency of the recommended development scenario generally increases, but the power of the recommended development scenario generally decreases. This pattern suggests a tradeoff between power and efficiency.

Maximal corrected relative efficiency is seen at $k = 3,4$.

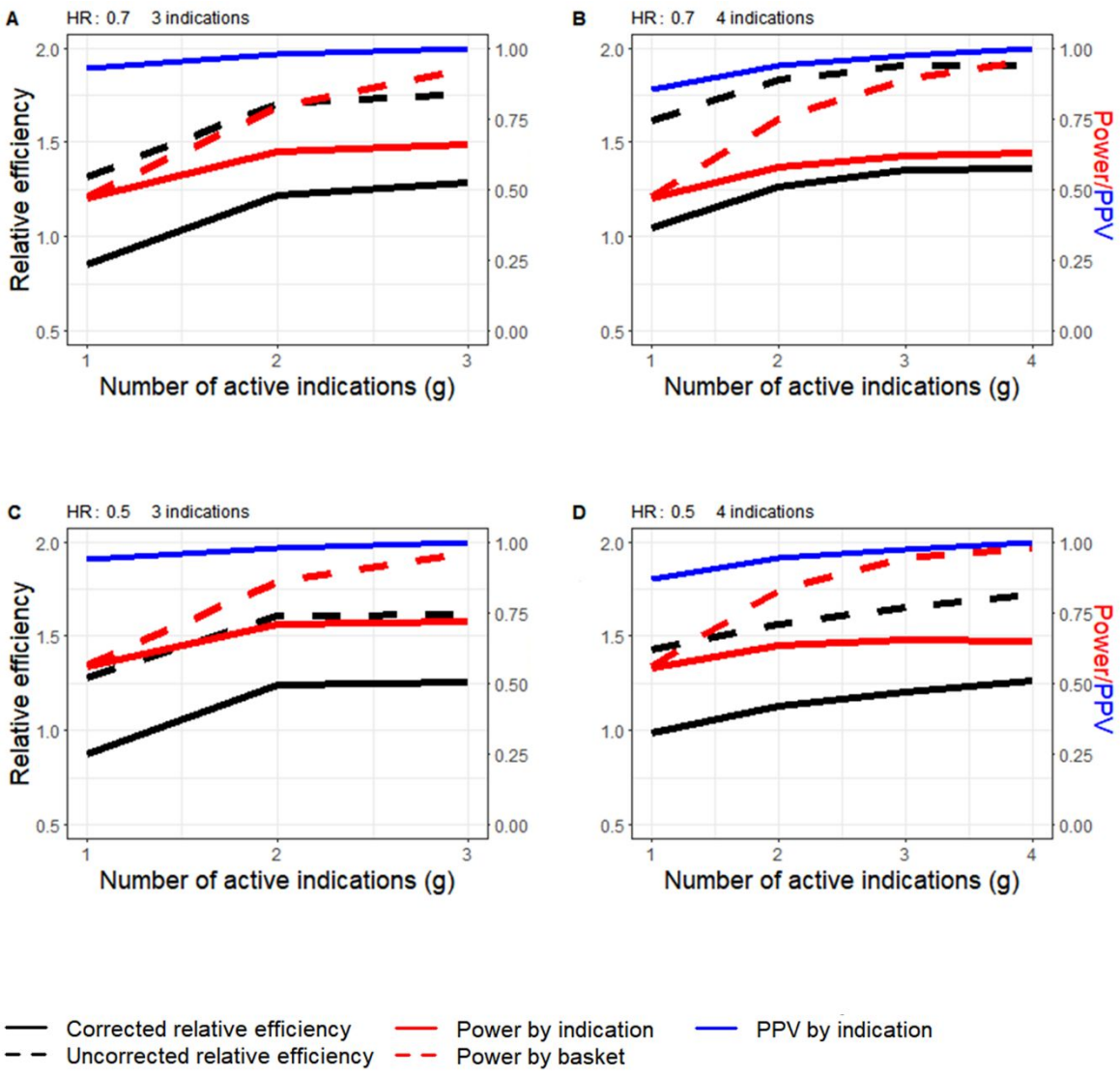


Figure 5. Recommended development approaches for (A) 3 indications with $HR = 0.7$ (Design 3, $kn = 409$, $\beta = 0.05$, $\alpha_t = 0.4$, $\alpha_{post} = 0.05$, $\alpha^* = 0.009$), (B) 4 indications with $HR = 0.7$ (Design 3,

$kn = 483$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0075$), (C) 3 indications with $HR = 0.5$ (Design 3, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.35$, $\alpha_{post} = 0.05$, $\alpha^* = 0.0089$), and (D) 4 indications with $HR = 0.5$ (Design 2, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0094$). The x-axis represents the number of active indications (indications in which the drug provides clinical benefit), the primary y-axis (left) represents the uncorrected/corrected relative efficiency, and the second y-axis (right) represents the power (red) by indication and by basket, and the positive predictive value by indication (blue).

Recommended development scenarios

We have previously emphasized that this design requires careful initial indication selection.[13,14] Users should strive to include only zero or one inactive indications in the study (Discussion). To determine recommended development scenarios, we considered these upside scenarios, specifically requiring power by indication greater than 60% when all indications are active, assuming that lower power might not be considered acceptable by some practitioners for a confirmatory study. We assumed that for the very modest therapeutic effect sizes considered herein, corresponding to effective but not transformational therapies, that a reduction in power from the traditional $80 - 90\%$ to 60% would be potentially applicable given

the marked increase in development efficiency offered, in settings where financial resources or available trial participants were limiting. In practice, the practitioner may use these simulation methods to set their own desired power cutoff, and higher power cutoffs may be possible for larger effect sizes. We simulate a range of values for k, β, D, α_t, and α_{post}. We note that α^* is not independently adjustable, but is calculated based on the other input parameters and the assumed hazard ratios using previous methods (Supplementary Material).[13] After implementing this filtering criterion for power in addition to the requirements for control of FWER, estimation bias, and confidence interval coverage at all values of g, only the scenarios with $k = 3, 4$ remain. Figure 5 summarizes the scenarios meeting this minimal power criterion with optimal upside corrected relative efficiency for $k = 3, 4$ and hazard ratio values of 0.5 and 0.7. Figure S1 presents the same information for hazard ratios 0.6 and 0.8. Power by basket increases from 50% to greater than 90% for any scenario as the number of active indications increases from 1 to k, while the power by indications increases to 63% -72% when all indications are active, and 62% -71% when there is one inactive indication. NPV by indication decreases from 91% at $g = 1$ to 71% and 55% for $g = k - 1$ for $k = 3$ and 4, respectively. PPV by indication increases from 86%

$(k = 4)$ -93% $(k = 3)$ to 98% as g increases from 1 to $g = k - 1$ for $k = 3, 4$, but remains stable whether the hazard ratio of positive indications increases or decreases. For the uncorrected relative efficiency, the relative efficiency is compared to a group of parallel independent studies at 90% power, yielding 90% average power overall. Scenarios with 4 indications exhibit about $43\% - 92\%$ efficiency improvement depending on g, $23\% - 78\%$ efficiency improvement for three indications. For the corrected relative efficiency, the basket scenarios are compared to k independent trials with the same power as achieved in the basket scenario, so that the average power is the same across the comparison. Corresponding ranges of corrected relative efficiency improvement are $-17\% - 29$ %, and $-1\% - 38\%$, for $k = 3, 4$, respectively. Note if most of the indications are inactive, the design is inefficient. When all indications are active, or only one indication is inactive, the uncorrected efficiency improvement ranges from 61% -78% and $65\% - 92\%$ for 3 and 4 indications, respectively, while the corrected improvement in efficiency ranges from $21\% - 29\%$ and $20\% - 38\%$ for 3 and 4 indications, respectively.

Optimal corrected relative efficiency scenarios

Without filtering for power, Figure 6 and S2 summarize scenarios with maximal corrected relative efficiency for $k = 3, 4, 5$ and different values of the hazard ratio. Scenarios with $k = 6$ have inferior performance (Supplemental Table S2) and are not recommended. In the discussion, we consider explanations for the optimal indication number range.

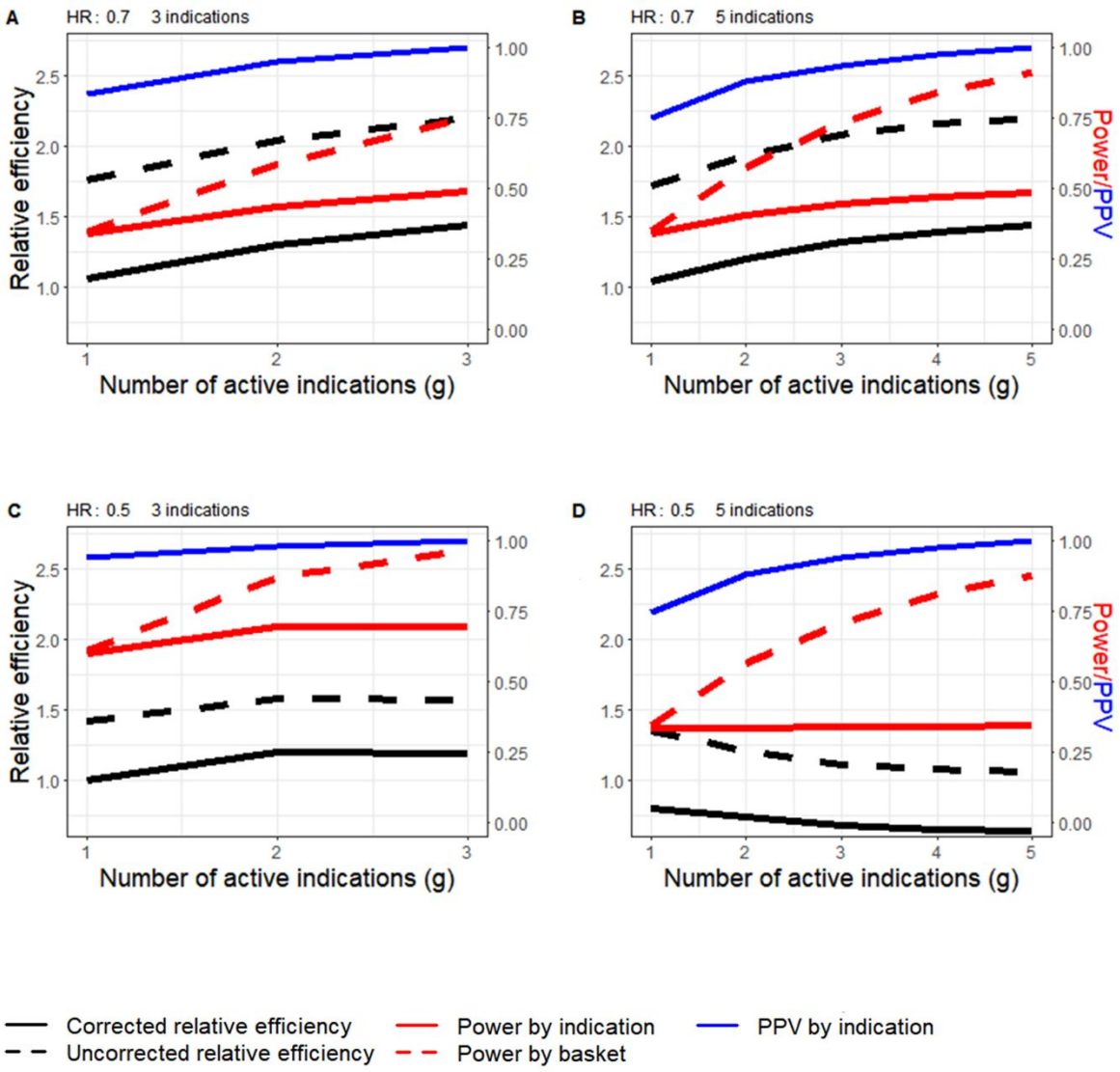


Figure 6. Cases with maximum corrected relative efficiency for (A) 3 indications with $HR = 0.7$ (Design 3, $kn = 247$, $\beta = 0.2$, $\alpha_t = 0.2$, $\alpha_{post} = 0.15$, $\alpha^* = 0.0101$), (B) 5 indications with $HR = 0.7$

(Design 3, $kn = 409$, $\beta = 0.05$, $\alpha_t = 0.15$, $\alpha_{post} = 0.15$, $\alpha^* = 0.0071$), (C) 3 indications with $HR = 0.5$ (Design 3, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.25$, $\alpha_{post} = 0.05$, $\alpha^* = 0.0093$), and (D) 5 indications with $HR = 0.5$ (Design 2, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.05$, $\alpha_{post} = 0.4$, $\alpha^* = 0.0278$). The x-axis represents the number of active indications (indications in which the drug provides clinical benefit), the primary y-axis (left) represents the uncorrected/corrected relative efficiency, and the second y-axis (right) represents the power (red) by indication and by basket, and the positive predictive value by indication (blue).

Comparing to the recommended development scenarios, the optimal corrected relative efficiency scenarios present higher efficiency, lower power, and lower NPV/PPV by indication. Power by basket ranges from $34\% - 98\%$, $26\% - 98\%$, and $21\% - 98\%$, depending on g, for $k = 3,4,5$, respectively. Corresponding power ranges for power by indication are $34\% - 71\%$, $25\% - 65\%$, and $20\% - 59\%$. PPV by indication ranges from $83\% - 100\%$, $71\% - 100\%$, and $55\% - 100\%$, for $k = 3,4,5$, respectively, which are lower than those of recommended scenarios. Uncorrected relative efficiency improvement ranges from $25\% - 125\%$, $47\% - 163\%$, and $6\% - 172\%$ depending on the value of g for $k = 3,4,5$ respectively. Corresponding ranges of corrected relative efficiency improvement are $-9\% - 47\%$, $-6\% - 67\%$, and $-36\% - 66\%$. When all indications are active or

only one indication is inactive, the uncorrected efficiency improvement ranges from $46\% - 125\%$, $65\% - 163\%$, and $6\% - 172\%$, for $k = 3, 4, 5$, respectively, while the corresponding corrected relative efficiency improvement ranges from $14\% - 47\%$, $20\% - 66\%$, and $-36\% - 66\%$. Removing the constraint on power increases relative efficiency gains, again indicative of the tradeoff between power and efficiency. In most cases, the basket trial design improves the efficiency, but this is not always the case (Figure 6D). This figure considers a case in which the potential level of heterogeneity is too great for the design to efficiently compensate for, in that there are a large number of indications (5) and active indications are postulated to be quite different from inactive ones (hazard ratio of 0.5 compared to 1.0). Controlling the Type I error by indication for all possible values of g in this situation forces a very stringent interim check, resulting also in the elimination of many true positives.

DISCUSSION

Confirmatory basket trials potentially provide remarkable improvements in drug development efficiency. With pooling, a fold-improvement in efficiency comparable to the number of indications is possible while retaining high power for the pool, and controlling alpha with respect to a global null hypothesis, in the randomized confirmatory basket design.[13, 14]

We studied a modification of the randomized confirmatory basket design[13, 14] in a more rigorous setting where control of FWER by indication subgroup is recommended, as may be needed for health authority approval in some instances.[15] We have further studied the most challenging scenario, i.e. a slowly maturing TTE endpoint without a highly predictive surrogate. Performance characteristics for innovative designs should be publicly disclosed in detail, including in challenging settings, but this is not always the case.[16] Under the conditions of the simulation, up to 92% improvement in relative efficiency, or 38% improvement corrected for reduced power by indication, is still possible, while controlling the FWER at a rate comparable to an equal number of parallel single-indication studies. Power of the pooled analysis ("power by basket") remains high, and estimation bias and confidence interval coverage may also be well controlled. Power by indication can be as high as $60\% - 73\%$, but declines if there are two or

more inactive indications in the basket. These results are possible only by using simulation to extensively optimize various design parameters, as described in this work. The results apply only to the conditions of the simulation. As there are an infinite number of cases to simulate, a sponsor would need to reach agreement with health authorities on the range of scenarios to be simulated, as has been articulated previously for complex innovative designs.[19]

We previously published a related design which controls Type I error *by basket* in the pooled data only[13, 14] and while a formal efficiency analysis was not performed, an application example suggested 200-500% improvement in efficiency. The current variant, that considers FWER by indication subgroup (see Figure 2 for the definition of control by indication subgroup as distinguished from control by basket), is less efficient than the original, but still substantially improved compared with independent parallel studies of each indication. Both the original and current designs contain one or more checkpoints that operate on individual indications, but also contain a checkpoint that operates on pooled data, and this checkpoint, the most stringent one, is far more efficient due to the pooling, which allows multiple indication subgroups to be considered for the sample size that would normally be required for one indication. In the present variant, the Type I error is actually

controlled *over the whole study by indication subgroup* at 0.025k, reflecting checkpoints some of which operate on pooled data, rather than individually at 0.025 per indication subgroup. Only in the example in which all indications are treated identically would the indication subgroups all be controlled at 0.025 individually.

The optimal range for indication number is narrow. For recommended drug development scenarios, which seek optimal corrected relative efficiency with a minimum constraint on power by indication, either 3 or 4 indications are optimal. If only corrected relative efficiency is optimized, one may also consider 5 indications. Too few indications and one does not get the benefit of a basket trial. Too many, and compensating for the large number of potential heterogeneity scenarios involved in controlling FWER becomes challenging. Interestingly, earlier work determining an optimal indication number in exploratory basket trials also recommended 3–5 indications despite a very different scenario and approach to optimization.[20]

Optimal stringency of filtering varies, but generally greater stringency is applied in post-individual tests than at interim. Higher stringency is perhaps better applied to more mature datasets. Optimal sample size adjustment (SSA) strategies varied.

Sometimes moderate SSA (D3) was better than aggressive SSA (D2). Aggressive SSA may optimize power at the expense of overall development efficiency, another example of a case where focusing on power leads to diminishing returns when considered from a portfolio perspective.[21] The optimal nominal power for the pooled analysis also varies according to the power/efficiency tradeoff. Higher power requires disproportionate investment, as previously shown for proof of concept studies.[2]

A study of ten oncology drugs found an average cost of $648 million for their clinical development, much of which spent in the resource-intensive confirmatory phase.[22] Expense and prolonged clinical development time delays availability of therapies, and contributes to their high cost and potentially to unequal treatment access, major public health issues. We believe the confirmatory basket design or modifications thereof could contribute to the solution of these problems. Moreover, in oncology niche indications, this design could make drug development economically feasible even in the absence of a transformational therapy. Transformational approaches such as immunotherapy must be further optimized, creating a broad universe of potential combination studies across populations sharing common characteristics such as high tumor mutation burdens. In this

example, immune checkpoint therapy could be combined in the experimental arm with another drug designed to improve its performance in multiple tumors. These therapy combinations may lead to important but incremental improvements that require a randomized TTE approach for confirmation. Potential applications are not limited to oncology. We are currently investigating the use of real world data of off-label use to design and simulate a basket trial in multiple autoimmune diseases in which rituximab is added to a standard of therapy arm consisting of steroid therapy.[23, 24]

We have previously suggested that the performance of randomized confirmatory basket trials depends on careful indication selection.[14] The importance of indication selection is even greater when control of FWER is recommended, as is documented in the results, in which performance improves with the proportion of indications that are active. The more inactive indications are present, the more stringent pruning and/or post check steps are required to reliably eliminate them and control the FWER by indication. Highly stringent pruning and/or post check steps run a greater risk of also eliminating some active indications, reducing power and efficiency. The more inactive indications are present, the greater the risk that they will dilute the signal in a pool containing positive indications.

These designs are therefore inappropriate for a collection of miscellaneous uncharacterized indications. Any proposed indication should, when feasible, be supported by preclinical studies, and Phase 2 clinical and biomarker data, ideally from randomized biomarker-guided Phase 2 studies.[14] Alternatively, it may be helpful to filter these indications in Phase 2 with one of several exploratory basket trial designs,[5-8] especially if the indications have enrollment challenges. Real world data/evidence, especially concerning off-label use, may be of value in screening potential indications.[23, 24] Ideally, one lead indication should have been confirmed with the biomarker in a biomarker-guided traditional Phase 3 study, and the basket trial would then be confirming supplemental indications using the same biomarker assay adjusted as needed for different tissue types.[14]

For the vast majority of therapies, a positive randomized Phase 2 proof of concept study is followed by a randomized Phase 3 confirmatory study. In analogous fashion, this design and its original predecessor[13, 14] are therefore expected to be applied for the majority of effective therapies, ideally supported by randomized Phase 2 data. In contrast, previous applications of confirmatory basket trials have been limited to a small number of exceptional therapies.

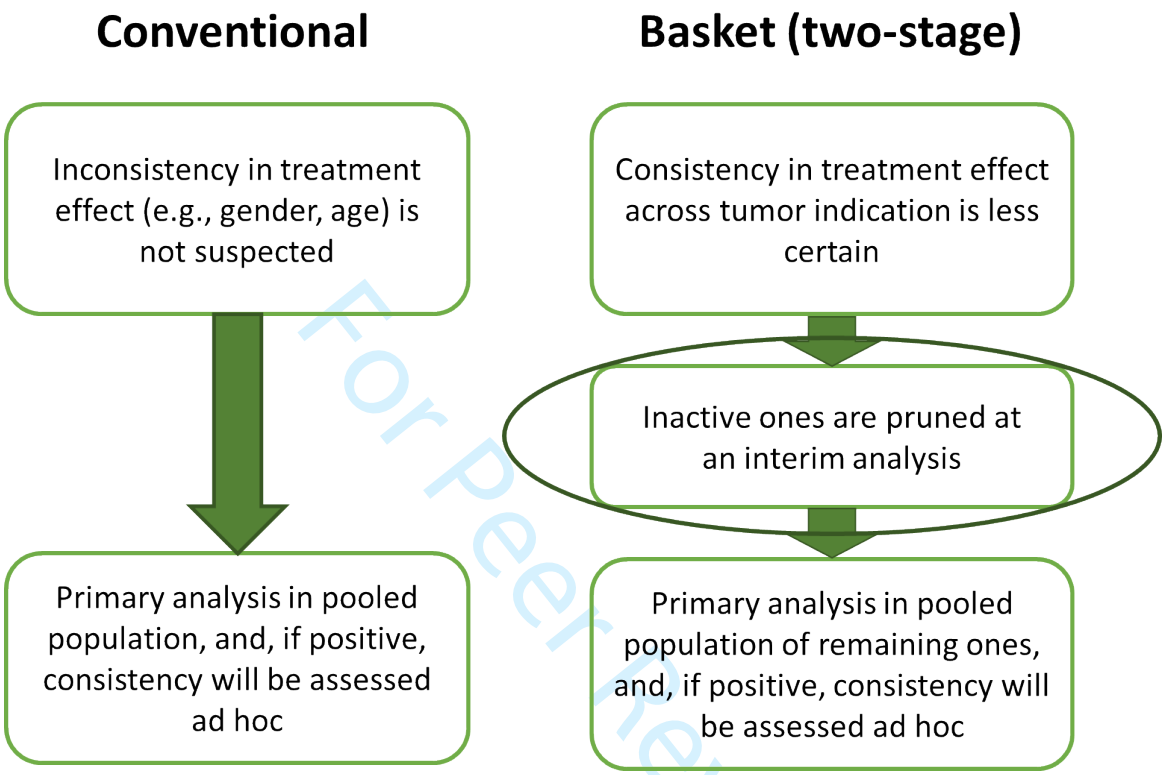


Figure 7. Parallels between a conventional Phase 3 study (left) and the randomized confirmatory basket trial design considered in previous work without strong control of the FWER by subgroup (right). Both designs formally test a hypothesis for a main group and do not formally test hypotheses involving subgroups. In the conventional design, organ site is the defining characteristic of the main group, and both known and unknown subgroups are present, the former perhaps undergoing more informal subgroup analyses. In the original randomized confirmatory basket trial design[13,14] the traditional organ site classification is only a known subgroup, and a

biomarker or other pathophysiologic feature defines the main group. Organ site classification is subjected to informal analysis only.

It is important to consider if and when strong control of FWER by indication subgroup should be recommended for a confirmatory basket trial with pooling, an area of debate which will likely evolve. In cases where control of Type I error by basket is permitted by health authorities, our original design[13, 14] should be used. However, should health authorities require control of Type I error by indication, the design in the current study would be suitable and still provide significantly improved development efficiency relative to conventional approaches. In both cases, the indication would be for a group of organ-specific indications sharing a common biomarker or pathophysiologic mechanism.

One might argue that Type I error control by basket should be adequate in many cases. A conventional Phase 3 study is really quite heterogeneous with respect to both known and unknown subgroups (Figure 7), and we do not control FWER in assessing the vast majority of these subgroups, employing other approaches.[25, 26] In a basket trial, we invert the usual classification: molecular subgroups are now indication-defining, and organ sites, formerly indication-defining, are now

subgroups. This alone may present a perceptual barrier to full acceptance of the concept. Collignon et al.[15] (an author group representing individuals with past or current associations with European health authorities) argue that homogeneity of outcomes may be difficult to interpret clinically since the populations are “different”, illustrating the unproven perception that differences in organ sites are more fundamental than the many other known and unknown differences between subpopulations. Nonetheless they also provide a definition of subgroup homogeneity in striking agreement with our thinking: “homogeneous if they share important clinical characteristics such that, in light of the available scientific evidence, the interpretation of treatment effect and the assessment of benefit/risk are meaningful for the overarching target population…” This comes down in the end to science and medicine, not statistics, and indeed the scientific justification for pooling must be robust if we are to forego FWER control by subgroup. We know that even in the classic case of an antagonist of a driver mutation in the b-raf gene, drug effectiveness still depends on organ site.[10] Increased understanding of how driver gene mutations interact with tissue-specific gene expression programs may be important. If we do not have a robust justification for pooling, control of FWER by indication will be necessary. Collignon et al.[15(SI)] evaluate our original

confirmatory basket design,[13,14] and our assertion that evidence supporting a consistent benefit/risk assessment across traditional indications should be provided at a level prospectively agreed with health authorities.[14] They state that availability of post-approval data would be important. This should become more practical as electronic health records systems improve. Informal interactions with health authorities further indicate that control of the FWER by indication is a controversial and evolving issue and may be required in some instances, and therefore there is a practical need for characterizing the performance of a randomized confirmatory basket design constrained to control the FWER by indication. The choice between our original design[13, 14] and the current design can thus be determined only based on discussion about required Type I error control with health authorities.

In ongoing research, we are considering important questions regarding the suitability of these designs for the confirmatory phase, in particular how to deal with differences between indications in endpoints and in safety issues. In future work, we will consider the effects of different control therapies, enrollment rates, and endpoint maturation rates. We are investigating real world data/evidence in indication screening and parameter estimation in simulations. Finally, current

performance may be improved further by application of Bayesian techniques previously devised for exploratory trials.[5-8, 27]

REFERENCES

1. Beckman RA, Clark J and Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature Reviews Drug Discovery* 2011; 10: 735-749.
2. Chen C and Beckman RA. Maximizing return on socioeconomic investment in Phase II Proof-of-Concept trials. *Clinical Cancer Research* 2014; 20: 1730-1734.
3. Woodcock J and LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine* 2017; 377: 62-70.
4. Antonijevic Z and Beckman RA. *Platform trial designs in drug development: Umbrella trials and basket trials*. Boca Raton: Chapman & Hall/CRC Biostatistics Series, Taylor & Francis Group, 2018.
5. Cunanan KM, Iasonos A, Shen R, et al. An efficient basket trial design. *Stat Med* 2017; 36: 1568-1579.
6. Berry SM, Broglio KR, Groshen S, et al. Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology trials. *Clinical Trials* 2013; 10: 720-734.
7. Chu Y and Yuan Y. A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials* 2018; 15: 149-158.
8. Simon R, Geyer S, Subramanian J, et al. The Bayesian basket design for genomic-variant driven phase II trials. *Semin Oncol* 2016; 43: 13-18.

9. Heinrich MC, Joensuu H, Demetri GD, et al. Phase II open-label study evaluating the activity of imatinib in treating life-threatening malignancies known to be associated with imatinib-sensitive tyrosine kinases. *Clin Cancer Research* 2008; 14: 2717-2725.
10. Hyman DM, Puzanov I, Subbiah V, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine* 2015; 373: 726-736.
11. Drilon A, Laetsch TW, Kummar S, et al. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. *New England Journal of Medicine* 2018; 378: 731-739.
12. Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine* 2015; 372: 2509-2520.
13. Chen C, Li N, Yuan S, et al. Statistical design and considerations of a phase 3 basket trial for simultaneous investigation of multiple tumor types in one study. *Statistics in Biopharmaceutical Research* 2016; 8: 248-257.
14. Beckman RA, Antonijevic Z, Kalamegham R, et al. Adaptive design for a confirmatory basket trial in multiple tumor types based on a putative predictive biomarker. *Clinical Pharmacology and Therapeutics* 2016; 100: 617-625.
15. Collignon O, Gartner C, Haidich AB, et al. Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trial. *Clinical Pharmacology and Therapeutics* 2020; 107: 1059-1067.
16. Cunanan KM, Iasonos A, Shen R, et al. Specifying the true- and false-positive rates in basket trials. *JCO Precision Oncology* 2017; 1:1-5.

17. Freter RR and Savageau MA. Proofreading systems of multiple stages for improved accuracy of biological discrimination. *J Theor Biol* 1980; 85: 95-123.
18. Beckman RA and Loeb LA. Multistage proofreading in DNA replication. *Quarterly Reviews of Biophysics* 1993; 26: 225-331.
19. U.S. Food and Drug Administration. Interacting with the FDA on Complex Innovative Trial Designs for drugs and biological products: Draft guidance for industry, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/interacting-fda-complex-innovative-trial-designs-drugs-and-biological-products (2019, accessed 30 June 2020).
20. Chen C, Deng Q, He L, et al. How many tumor indications should be initially screened in clinical development of next generation immunotherapies? *Contemporary Clinical Trials* 2017; 59: 113-117.
21. Antonijevic Z. The impact of adaptive design on portfolio optimization. *Therapeutic Innovation and Regulatory Science* 2016; 5: 615-619.
22. Prasad V and Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Intern Med* 2017; 177(11):1569-1575.
23. Guinn D, Ren Y, Wilhelm EE, et al. Utilizing real-world data to inform a confirmatory basket trial design- studying use of rituximab in autoimmune diseases. *Value in Health* 2018; 21: S229-30.

24. Guinn D, Madhavan S and Beckman RA. *Harnessing Real-World Data to Inform Platform Trial Design.* Platform Trial Designs in Drug Development: Umbrella Trials and Basket Trials, 2018, p.55.
25. European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials, https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf (2019, accessed 30 June 2020).
26. ICH Expert Working Group. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9. *International Conference on Harmonisation (ICH)*, 1998.
27. Zhou H, Chen C, Sun L, et al. Bayesian optimal phase II clinical trial design with time to event endpoint. *Pharmaceutical Statistics*. Epub ahead of print 10 June 2020. https://doi.org/10.1002/pst.2030.

ACKNOWLEDGEMENTS

We thank Drs. Zoran Antonijevic, Lisa LaVange, Peter Mesenbrink, Martin Posch, William Rosenberger, and Alex Sverdlov for careful and insightful review of the manuscript.

DISCLAIMER

Dr. Daphne Guinn contributed to this article in her personal capacity. The views expressed are her own and do not necessarily represent the views of the Food & Drug Administration or the United States Government.

SUPPLEMENTAL METHODS

The adjusted nominal level α* in the pooled analysis[13, 14]

The pooled analysis aims to examine that there is treatment effect in at least one tumor indication. This analysis combined with the previous pruning and sample size readjustment steps is controlled at type I error of 0.025 for the global null hypothesis that all indications are inactive using methods from reference 13. Note that the definition of Type I error of the entire trial in reference 13 is different from the FWER in this study. Specifically, in reference 13, let Y_{i1} be the standardized test statistics based on the endpoint used for pruning at the interim analysis, and Y_{i2} be the standardized test statistics based on the endpoint for pooling for the i-th tumor indication at the final analysis ($i = 1,...,k$). Suppose that m tumor indications are included in the pooled analysis ($m \geq 1$). Let V_m be the corresponding standardized test statistics pooled from Y_{i2}, which can be written as $\left(\sum_{i=1}^{m} Y_{i2}\right)/\sqrt{m}$. Under the null hypothesis H_0 that there is no treatment effect in any of the tumor indications, the probability of incorrectly declaring activity in a basket, denoted as α, is

$$\alpha = \sum_{m=1}^{k} c(k,m) Q_0(\alpha^* \mid \alpha_t, m)$$

where $c(k,m) = k!/((k-m)!m!)$ is the number of choices for selection of m tumor indications from k, and $Q_0(\alpha^* | \alpha_t, m)$ is the probability of V_m being statistically significant at the adjusted α^* level given m out of k indications in the pool, formulated as

$$Q_0(\alpha^* | \alpha_t, m) = P_{H_0}\left(\cap \{Y_{i1} > Z_{1-\alpha_t}; i = 1,...,m\}, \cap \{Y_{i1} < Z_{1-\alpha_t}; i = m+1,...,k\}, V_m > Z_{1-\alpha^*} \right)$$

Assuming that $\{Y_{i1}; i = 1,...,k\}$ are ***i.i.d.***, under the global null hypothesis we have

$$Q_0(\alpha^* | \alpha_t, m) = P_{H_0}\left(\cap \{Y_{i1} > Z_{1-\alpha_t}; i = 1,...,m\}, V_m > Z_{1-\alpha^*} \right)(1-\alpha_t)^{(k-m)}$$

Consider $\{Y_{11}, ..., Y_{m1}, V_m\}$ following a multivariate normal distribution $(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{pmatrix} 1 & & 0 & corr(Y_{11},V_m) \\ & \ddots & & \vdots \\ 0 & & 1 & corr(Y_{m1},V_m) \\ corr(Y_{11},V_m) & ... & corr(Y_{m1},V_m) & 1 \end{pmatrix}$. Setting $\alpha = 0.025$, the adjusted level α^*, at which the pooled analysis is nominally set, can be solved based on the correlation between Y_{i1} and V_m, $corr(Y_{i1},V_m)$[13].

In this study, we consider the following three sample size adjustment strategies[13] and the corresponding $corr(Y_{i1},V_m)$ can be calculated as:

1. Design one (D1): The sample size for each tumor indication is fixed upfront at *n* as planned. After pruning, the sample size will be less than or equal to the originally planned kn. Under D1, $corr(Y_{i1},V_m) = \sqrt{t/m}$. This strategy corresponds to no sample size adjustment.

2. Design two (D2): The sample size for each tumor indication will increase after the interim analysis so that the total sample size in the pooled analysis remains kn, which is greater than the sample size under D1 after pruning. Under D2, $corr(Y_{i1},V_m) = \sqrt{t/k}$. This is the most aggressive sample size adjustment strategy.

3. Design three (D3): The sample size for each tumor indication will increase after the interim analysis so that the total sample size in the overall study remains kn. Thus, the total sample size in the pooled analysis is greater than the sample size under D1 but smaller than the sample size under D2. Under D3, $corr(Y_{i1},V_m) = \sqrt{t/(mt + k(1-t))}$.

SUPPLEMENTAL TABLES

Table S1. Glossary of terms for simulation study

Design parameters	Value(s) in simulation study	Descriptions
The state of nature		
g	$0,...,k$	The number of active indications at the beginning
θ_i	At null value $\theta_0 = 1$, At active value $\theta_1 = 0.5,0.6,0.7,0.8$.	The hazard ratio of experimental arm vs. control arm for the definitive endpoint of each indication
Study design input parameters		
k	$3,4,5,6$	The number of tumor indications at the beginning
D	D1, D2, D3	Sample size adjustment strategies
t	0.5	A common information time for the interim analysis
α_t	$0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4$	A common bar for the interim analysis to prune inactive indications
α	0.025	The false positive rate for pooled analysis together with the pruning step with respect to the global null hypothesis, after inflation from the nominal value α^*
β	$0.025,0.05,0.1,0.2$	False negative rate for the pooled analysis

α_{post}	0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4	A common bar for the post -individual test, which is varied independently of α_t
α_{ref}	0.025	False positive rate for the reference design
$\beta_{uncorrected}$	0.1	False negative rate for the uncorrected reference design
S	10000	The number of simulated replications
Calculated parameters		
n	$n = \frac{4(Z_{1-\alpha} + Z_{1-\beta})^2}{k(log\theta)^2}$	Planned sample size for each tumor indication
α^*	Calculated by numerically solving the equation $\sum_{i=1}^{k} c(k,m) Q_0(\alpha^*, \alpha_t, m) = \alpha$	The adjusted nominal level α^*, at which the pooled analysis is nominally set to control the false positive rate α for the pooled analysis combined with the pruning step.
$\beta_{corrected}$	1 – power by indication corresponding to the given input parameters as determined by simulation	False negative rate for the corrected reference design
Simulated parameters	Descriptions	
m	The number of tumor indications included in the pooled analysis	
n_{i1}	Sample size for each tumor indication at the interim analysis (nt)	
n_{i2}	Sample size for each tumor indication at the pooled analysis	
Y_{i1}	The standardized test statistics for the i-th tumor indication at the interim analysis	
Y_{i2}	The standardized test statistics for the i-th tumor indication at the final analysis	
V_m	The standardized test statistics pooled from Y_{i2}	
d	The number of false positive indications passing the final individual tests	
j	The number of true positive indications passing the final individual tests	

Outcome measurements	Estimates in simulation study	Descriptions
α_{net} (g)	$\frac{1}{S}\sum_{s=1}^{S} I\left(V_m^{(s)} > Z_{1-\alpha^{(s)}}\right) I\left(d^{(s)} > 0\right)$	The probability of the basket trial passing the pooled test and at least one false positive indication passing the post-individual test for a given value of g.
Family-wise error rate (FWER)	$max\{\alpha_{net}(g): g = 0,..., k\}$	For any g ($g = 0,..., k$), the maximum probability of the basket trial passing the pooled test and at least one false positive indication passing the post-individual test
Power by indication	$\frac{1}{S}\sum_{s=1}^{S} \frac{j_{new}^{(s)}}{g} I\left(V_m^{(s)} > Z_{1-\alpha^{(s)}}\right)$	The proportion of true positive indications that pass the post- individual test (requires that the pooled test passed)
Power by basket	$\frac{1}{S}\sum_{s=1}^{S} [I(g > 0) I\left(V_m^{(s)} > Z_{1-\alpha^{(s)}}\right)]$	The probability of an active basket (one that contains at least one active indication) passing the pooled analysis
Sample size	$(k - m)n_{i1} + mn_{i2}$	The sample size of a basket trial
Efficiency	$\frac{power \times g}{\frac{1}{S}\sum_{s=1}^{S}\left(\sum_{i=1}^{m^{(s)}} n_{i2}^{(s)} + \sum_{i=m^{(s)}+1}^{k} n_{i1}^{(s)}\right)} = \frac{\sum_{s=1}^{S} j_{new}^{(s)} I\left(V_m^{(s)} > Z_{1-\alpha^{(s)}}\right)}{\sum_{s=1}^{S}\left(\sum_{i=1}^{m^{(s)}} n_{i2}^{(s)} + \sum_{i=m^{(s)}+1}^{k} n_{i1}^{(s)}\right)}$	The ratio of average number of active indications that passed the post-individual tests divided by the average sample size
Uncorrected reference efficiency	$\frac{g(1-\beta_{uncorrected})(log\theta_1)^2}{4k\left(Z_{1-\alpha_{ref}} - Z_{1-\beta_{uncorrected}}\right)^2}$	The ratio of estimated number of active indications divided by the pre-defined total sample

		size in the reference study powered at 90%.
Uncorrected relative efficiency	Efficiency/Uncorrected reference efficiency	The ratio of efficiency and uncorrected reference efficiency
Corrected reference efficiency	$\frac{g(1-\beta_{corrected})(log\theta_1)^2}{4k(Z_{1-\alpha_{ref}}-Z_{1-\beta_{corrected}})^2}$	The ratio of estimated number of true positive indications divided by the pre-defined total sample size in the reference study at the same power by indication observed in the simulation in the basket trial with corresponding parameters, investigating the same indications.
Negative predictive value (NPV) by indications	$\frac{1}{S}\sum_{s=1}^{S}\frac{k-g-d^{(s)}}{k-j^{(s)}-d^{(s)}}$	The proportion of true negative indications among all negative indications not passing the interim test, the pooled test, or the post individual test
Negative predictive value (NPV) by basket	$\frac{1}{S}\sum_{s=1}^{S}I(g=0)I(j^{(s)}+d^{(s)}=0)$	The proportion of true negative baskets among all negative baskets without any indication passing the interim test, the pooled test, or the post individual test
Positive predictive value (PPV) by indications	$\frac{1}{S}\sum_{s=1}^{S}\frac{j^{(s)}}{j^{(s)}+d^{(s)}}$	The proportion of true positive indications among all positive indications passing the interim test, the pooled test, and the post individual test
Positive predictive	$\frac{1}{S}\sum_{s=1}^{S}I(g>0)I(j^{(s)}+d^{(s)}>0)$	The proportion of true positive baskets among all positive baskets with at

value (PPV) by basket		least one indication passing the interim test, the pooled test, and the post individual test
Corrected relative efficiency	Efficiency/Corrected reference efficiency	The ratio of efficiency and corrected reference efficiency
Coverage for hazard ratio (HR)	$\frac{1}{S}\sum_{s=1}^{S}[\sum_{i=1}^{m^{(s)}} I\left(Y_{i2}^{(s)} > Z_{1-\alpha_{post}}\right) I(\lvert Y_{i2}^{(s)} + \log(HR$	The probability that the estimated 95% CI for HR covers the true HR given the individual test passed
Bias of estimated HR	$\frac{1}{S}\sum_{s=1}^{S}\{\left[\frac{1}{m^{(s)}}\sum_{i=1}^{m^{(s)}} \exp\left(-Y_{i2}^{(s)}\sqrt{\frac{4}{n_{i2}^{(s)}}}\right) I\left(Y_{i2}^{(s)} > Z_{1-$	The relative difference in the true HR and the estimated HR, defined as the ratio of estimated average HR and true pooled HR for those indications that pass the individual tests minus 1.

Table S2. Simulation results. Each row summarizes the results of 10000 simulations for a given scenario with input parameters: hazard ratio (HR), number of indications (k), β, α_t (α_t), α_{post} (α_post), number of active indications (g), and sample size adjustment strategies. The outcome measurements include: α_{net}(α_net), power by indication, power by basket, mean sample size over simulations, efficiency, 95% CI coverage for HR, bias, uncorrected reference efficiency, corrected reference efficiency, uncorrected relative efficiency, and corrected relative efficiency.

SUPPLEMENTAL FIGURES

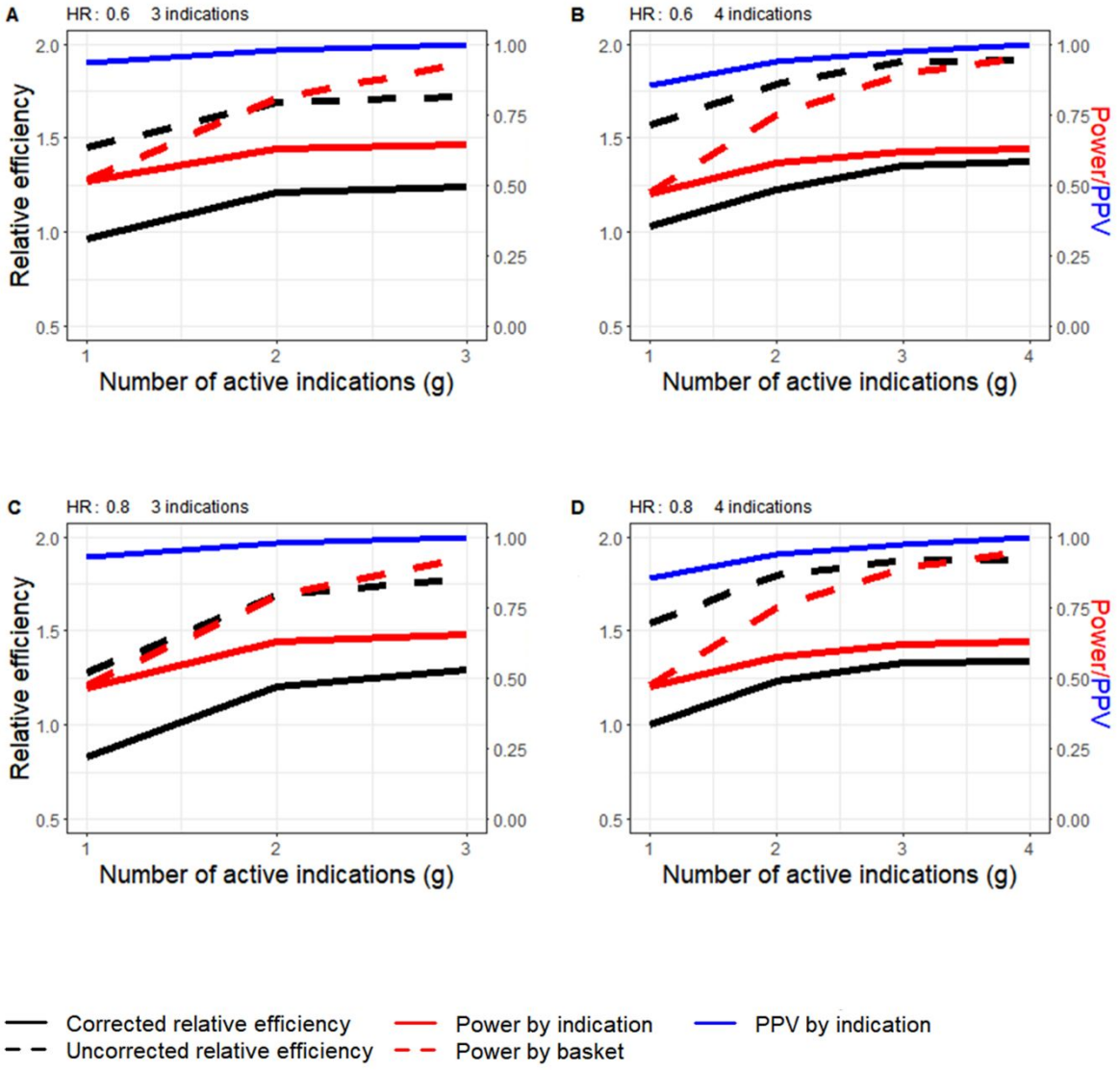


Figure S1. Recommended development approaches for (A) 3 indications with $HR = 0.6$ (Design 3, $kn = 199$, $\beta = 0.05$, $\alpha_t = 0.3$, $\alpha_{post} = 0.05$, $\alpha^* = 0.009$), (B) 4 indications with $HR = 0.6$ (Design 3, $kn = 236$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0075$), (C) 3 indications with $HR = 0.8$ (Design 3, $kn = 1044$, $\beta = 0.05$, $\alpha_t = 0.4$, $\alpha_{post} = 0.05$, $\alpha^* = 0.009$), and (D) 4 indications with $HR = 0.8$ (Design 3, $kn = 1234$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.01$, $\alpha^* = 0.0075$). The x-axis represents the number of active indications (indications in which the drug provides clinical benefit), the primary y-axis (left) represents the uncorrected/corrected relative efficiency, and the second y-axis (right) represents the power (red) by indication and by basket, and the positive predictive value by indication (blue).

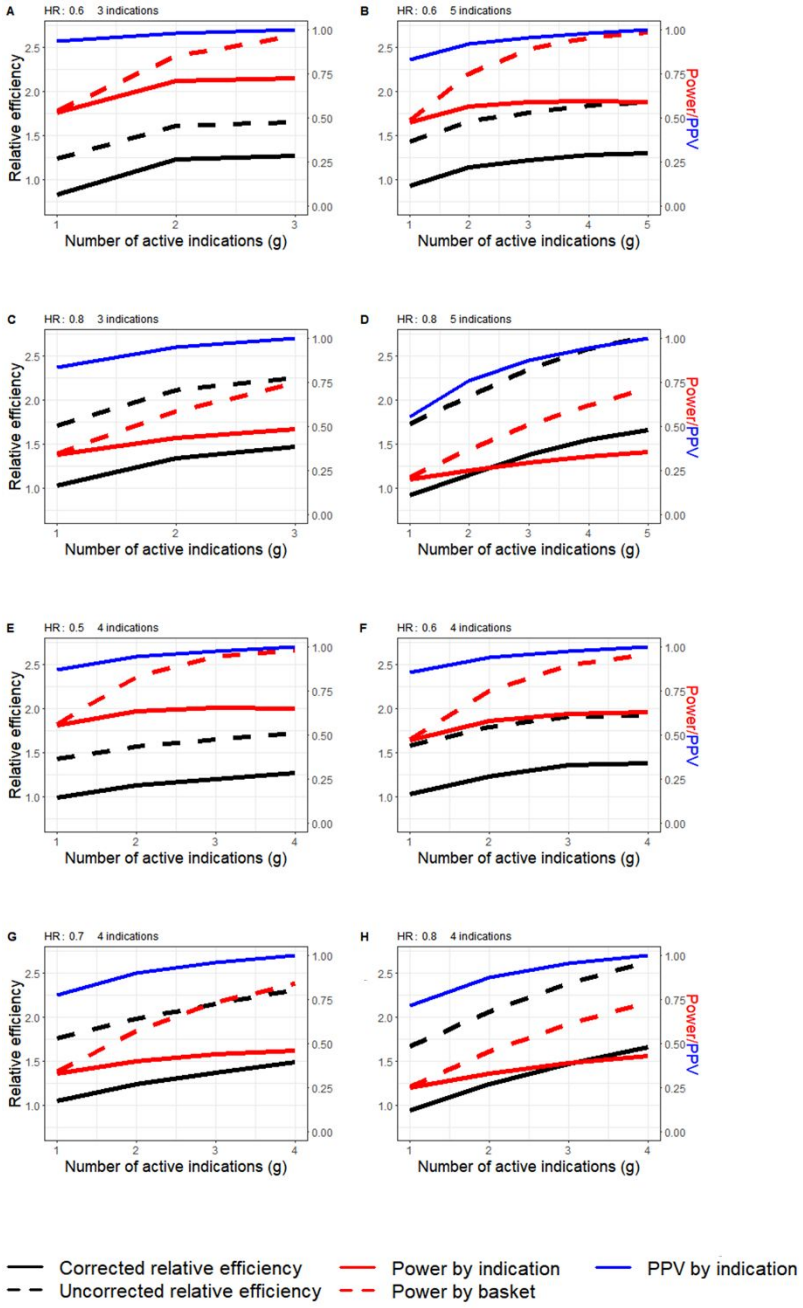

A
HR: 0.6 3 indications
B
HR: 0.6 5 indications
C
HR: 0.8 3 indications
D
HR: 0.8 5 indications
E
HR: 0.5 4 indications
F
HR: 0.6 4 indications
G
HR: 0.7 4 indications
H
HR: 0.8 4 indications
Relative efficiency
Power/PPV
Number of active indications (g)
Corrected relative efficiency
Uncorrected relative efficiency
Power by indication
Power by basket
PPV by indication

Figure S2. Cases with maximum corrected relative efficiency for (A) 3 indications with $HR = 0.6$ (Design 3, $kn = 236$, $\beta = 0.025$, $\alpha_t = 0.4$, $\alpha_{post} = 0.05$, $\alpha^* = 0.009$), (B) 5 indications with $HR = 0.6$ (Design 2, $kn = 236$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0078$), (C) 3 indications with $HR = 0.8$ (Design 3, $kn = 631$, $\beta = 0.2$, $\alpha_t = 0.2$, $\alpha_{post} = 0.15$, $\alpha^* = 0.0101$), (D) 5 indications with $HR = 0.8$ (Design 3, $kn = 631$, $\beta = 0.2$, $\alpha_t = 0.15$, $\alpha_{post} = 0.4$, $\alpha^* = 0.0071$), (E) 4 indications with $HR = 0.5$ (Design 2, $kn = 128$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0094$), (F) 4 indications with $HR = 0.6$ (Design 3, $kn = 236$, $\beta = 0.025$, $\alpha_t = 0.2$, $\alpha_{post} = 0.1$, $\alpha^* = 0.0075$), (G) 4 indications with $HR = 0.7$ (Design 2, $kn = 247$, $\beta = 0.2$, $\alpha_t = 0.2$, $\alpha_{post} = 0.15$, $\alpha^* = 0.0094$), and (H) 4 indications with $HR = 0.8$ (Design 3, $kn = 631$, $\beta = 0.2$, $\alpha_t = 0.2$, $\alpha_{post} = 0.2$, $\alpha^* = 0.0075$). The x-axis represents the number of active indications (indications in which the drug provides clinical benefit), the primary y-axis (left) represents the uncorrected/corrected relative efficiency, and the second y-axis (right) represents the power (red) by indication and by basket, and the positive predictive value by indication (blue).

SUPPLEMENTAL R CODES

```
#Calculate alpha* given alpha_t, information time, and the number of indications.
#Reference: Cong Chen, Xiaoyun (Nicole) Li, Shuai Yuan, Zoran Antonijevic,
#Rasika Kalamegham & Robert A. Beckman(2016) Paper: Statistical Design and
#Considerations of a Phase 3 Basket Trial for Simultaneous Investigation of
#Multiple Tumor Types in One Study, Statistics in  Biopharmaceutical Research,
#8:3, 248-257 DOI: 10.1080/19466315.2016.1193044. Function "mf2" is the original
#function from Chen's online code, it is used to express formula (3) in
#Chen-Beckman paper.

mf2 <- function(alphastar, alphat, t, m, k, d, Rho_for_endpoints=1) {
  # test1, test2, and test3 denotes correlation matrices for D1, D2, and D3.
  test <- matrix(0, ncol <- m + 1, nrow <- m + 1) # D1
  low <- rep(qnorm(1 - alphat), m + 1)
  low[m + 1] <- qnorm(1 - alphastar)
  up <- rep(Inf, m + 1)
  diag(test) <- 1

  for (i in 1:m){
    test[i, m + 1] <- test[m + 1, i] <- switch (d,Rho_for_endpoints * sqrt(t / m),
```

```
                                    Rho_for_endpoints * sqrt(t / k),
                                    Rho_for_endpoints * sqrt(t / (k * (1 - t) + m * t)))
}

  # joint_probability1, joint_probability2, and joint_probability3 denotes the
  # joint probability in equation (3) in Chen-Beckman paper for D1, D2, and D3.
  joint_probability <- pmvnorm(lower <- low,upper <- up,mean <- rep(0, m + 1),corr
<- test)[1]

  return(joint_probability)
}

# Function "type2" is the original function from Chen's online code, it is used
# to calculate alpha* by equation (3) in Chen-Beckman paper.
type2 <- function(alphastar, alphat, t, k, d, Rho_for_endpoints = .5) {
  # joint_probability denotes the joint probability in equation(3) in Chen-Beckman
paper.
  joint_probability <- 0

  for (i in 1:k)     joint_probability <- joint_probability + factorial(k) / (factorial(i) *
factorial(k - i)) *
      (1-alphat)^(k-i)*mf2(alphastar=alphastar, alphat=alphat, t=t, m=i, k=k,
d=d,Rho_for_endpoints=Rho_for_endpoints)
```

```
  return(joint_probability - 0.025)
}

#"Simulation" function is denoted to calculate Type I error and power
# D denotes D1, D2, and D3, alpha_t denotes Type I error in interim stage,
# alpha_tt denotes Type I error after final stage for the post-trial test.
Simulation=function(alpha_t, alpha_tt, g, k, t = 0.5, design=c(D1=1,D2=2,D3=3),
                    Rho_for_endpoints,hr, delta=-log(hr), n, simulation_times=10000) {

  #set seed before simulation is to guarantee the results are reproductable.
  set.seed(123)
  # print(dummy_indication)
  #t denotes information time
  dummy_indication= sample(c(rep(1, g), rep(0, k-g)))
  delta=delta * dummy_indication; hr = exp(-delta)
  # print(delta);print(hr)
  #alphastar denotes alpha* under D1, D2 and D3 by Chen-Beckman's formula (2),
the
  #function of "uniroot" is from Chen's online code.
  alphastar <- lapply(design,function(dd)
    uniroot(type2, c(0, 1), alphat=alpha_t, t=t, k=k,
d=dd,Rho_for_endpoints=Rho_for_endpoints)$root)
```

```
  mean_n <- delta * sqrt(n * t / 4)

  ###### Simulate result in before pruning

  x1 <- t(sapply(1:k,function(ln) if(dummy_indication[ln]==1)
{rnorm(simulation_times,mean=mean_n[ln], sd = 1)

    }else {rnorm(simulation_times, mean = 0, sd = 1)}))

  passed_pruning <- (x1 > qnorm(1 - alpha_t))


  #m denotes the number of remained indications after pruning

  m=colSums(passed_pruning)


  # Calculate sample size after pruning based on adjustment strategies.

  sample_size <- lapply(design,function(dd) apply(passed_pruning,2,
function(pptmp)

    switch (dd,n * pptmp, ceiling(k * n / ifelse(sum(pptmp)>0,sum(pptmp),Inf)) *
pptmp,

            ceiling((n * t + k * n * (1 - t) / ifelse(sum(pptmp)>0,sum(pptmp),Inf))) *
pptmp)))


  ## calculate total sample size in the trial

  total_sample_size <- lapply(design,function(dd) apply(passed_pruning, 2,
function(l)

    switch (dd,

            sum(n*l)+sum(n*t*(1-l)),
```

```
        ifelse(sum(l)>0,sum(n)+sum(n*t*(1-l)),sum(n*t*(1-l))),
        ifelse(sum(l)>0,sum(n),sum(n*t*(1-l)))
    )))


  # Calculate correlation between Yi1 and Yi2 if not all indications are pruned after
pruning step.
  rho_between_standardized_test_statistics <- lapply(design,function(dd)
sapply(m, function(mi)
    ifelse(mi>0, switch(dd,
                  Rho_for_endpoints * sqrt(t),
                  Rho_for_endpoints * sqrt(t * mi / k),
                  Rho_for_endpoints * sqrt(t / (t + k * (1 - t) / mi))
    ), 0)))


  # Generate Yi2 based on Yi1, corr(Yi1, Yi2), and adjusted sample size.
  #mean in the interim stage, mu1=sqrt(n*t/4)*delta*dummy_indication
  #representing means in the interim stage
  mu1 <- sqrt(n * t / 4) * delta * dummy_indication


  #mean in the final stage, mu2=sqrt(sample_size*1/4)*delta*dummy_indication
  #representing means in the final stage
```

```
  mu2 <- lapply(design, function(dd) (sqrt(sample_size[[dd]] / 4) * delta *
dummy_indication))


  #To generate Yi2, variance of Yi2 is sqrt((1-corr(Yi1,Yi2)^2))*s2^2,s2=1 in

  #our case(standardized normal distribution), given one of D1, D2, or D3.

  sd2 <- lapply(design, function(dd) sqrt((1 -
rho_between_standardized_test_statistics[[dd]] ^ 2)))


  #To generate Yi2, mean of Yi2(denoted as

  #mean_x2)=mu2+(s2/s1)*corr(Yi1,Yi2)*(Yi1-mu1),s1=s2=1 in our

  #case(standardized normal distribution) given one of D1, D2, or D3.

  mean_x2 <- lapply(design, function(dd)  mu2[[dd]] + (rep(1,k) %o%
rho_between_standardized_test_statistics[[dd]]) *
                    (x1 - mu1))


  #Generate Yi2 based on mean and variance for 3 indications, given one of D1,

  #D2, or D3.

  x2 <- lapply(design, function(dd) sapply(1:simulation_times, function(nsim)
    rnorm(mean_x2[[dd]][,nsim], mean_x2[[dd]][,nsim], sd = sd2[[dd]][nsim])))


  #Calculate V_m statistics, denotes as the sum of Yi2 for those indications

  #passed pruning at interim stage, divided by square root of number of

  #indications remained after pruning
```

```
vm <- lapply(design, function(dd) sapply(1:simulation_times, function(nsim)
  ifelse(m[nsim]>0,sum(x2[[dd]][passed_pruning[,nsim],nsim]) / sqrt(m[nsim]),0)))

# p_value_for_final_stage_testing denotes p-values for each simulation
#time to compare with alphastar later
  p_value_for_final_stage_testing <- lapply(vm, pnorm)

#If pooled indications are positive, do a post-check on individual
#indications at alpha=post_alpha_t, and discard indications that do not
#achieve statistical significance.
#passed_pruning_post_trial denotes indications that passed pruning and pass
#post-trial test
passed_pruning_post_trial <-lapply(design, function(dd)
  (x1 > qnorm(1 - alpha_t) & x2[[dd]] > qnorm(1 - alpha_tt)))

# calculate the coverage rate of 95% CI for individual indications and
# pooled indications.
# coverage of individual approved indications
coveragebias=lapply(design, function(dd)
sapply(1:simulation_times,function(nsim) {
```

```
pppttmp=passed_pruning_post_trial[[dd]][,nsim];wttmp=n[pppttmp]/sum(n[pppttmp]
)
   pppttmp.tp=(passed_pruning_post_trial[[dd]][,nsim]) & (dummy_indication==1)
   wttmp.tp=n[pppttmp]/sum(n[pppttmp])

   hrtmp=exp(- x2[[dd]][,nsim] * sqrt(4 / sample_size[[dd]][,nsim]))[pppttmp]
   hrtmp.tp=exp(- x2[[dd]][,nsim] * sqrt(4 / sample_size[[dd]][,nsim]))[pppttmp.tp]

   c(individual=ifelse(sum(pppttmp)>0,
                  sum(abs(mu2[[dd]][pppttmp,nsim]-x2[[dd]][pppttmp,nsim])<
                     qnorm(p = 0.025, lower.tail = F))/sum(pppttmp),NA),

     pooled=ifelse(sum(pppttmp)>0,
                abs(sum(x2[[dd]][pppttmp,nsim]) / sqrt(sum(pppttmp))+

log(sum(hr[pppttmp]*wttmp))*sqrt(sum(sample_size[[dd]][pppttmp,nsim]) / 4))<
                  qnorm(p = 0.025, lower.tail = F),NA),

     bias1=ifelse(sum(pppttmp)>0,mean(hrtmp)/ mean(hr[pppttmp])-1,NA),
     bias2=ifelse(sum(pppttmp)>0,sum(hrtmp * wttmp) / sum(hr[pppttmp] * wttmp)-
1,NA),
```

```
    indiv.tp=ifelse(sum(pppttmp.tp)>0,
                    sum(abs(mu2[[dd]][pppttmp.tp,nsim]-x2[[dd]][pppttmp.tp,nsim])<
                        qnorm(p = 0.025, lower.tail = F))/sum(pppttmp.tp),NA),
    pooled.tp=ifelse(sum(pppttmp.tp)>0,
                     abs(sum(x2[[dd]][pppttmp.tp,nsim]) / sqrt(sum(pppttmp.tp))+
log(sum(hr[pppttmp.tp]*wttmp.tp))*sqrt(sum(sample_size[[dd]][pppttmp.tp,nsim]) /
4))<
                     qnorm(p = 0.025, lower.tail = F),NA),
    bias1.tp=ifelse(sum(pppttmp.tp)>0, mean(hrtmp.tp)/ mean(hr[pppttmp.tp])-
1,NA),
    bias2.tp=ifelse(sum(pppttmp.tp)>0,sum(hrtmp.tp * wttmp.tp) /
sum(hr[pppttmp.tp] * wttmp.tp)-1,NA)
   )}))


 #Record tp, fp in each simulation time after interim stage
 #tp denotes the number of active remained indication after pruning, fp denotes
the number of remained
 #indication after pruning.


 fp= lapply(design,function(dd) colSums(dummy_indication==0 &
passed_pruning_post_trial[[dd]]==1))


 #j, denotes the number of active indications remained after pruning and pass
```

```
  #post-trial test
  #(if we don't do post trial test, we change passed_pruning_post_trial to
  #passed_pruning in this line.)
  tp=lapply(design, function(dd)  colSums(dummy_indication==1 &
passed_pruning_post_trial[[dd]]==1))


  # Use the formula of Type I error and powers to get simulation results.


  final_pooled_test=lapply(design, function(dd)
    m > 0 & p_value_for_final_stage_testing[[dd]] > (1 - alphastar[[dd]]))


  fptp <-lapply(design, function(dd)
    c(type_I_error=sum(final_pooled_test[[dd]] & fp[[dd]] > 0) / simulation_times,
      power1=ifelse(g>0,sum(tp[[dd]][final_pooled_test[[dd]]]) / (g *
simulation_times),0),
      power2=ifelse(g>0,sum(final_pooled_test[[dd]])/simulation_times,0)))


  ## Average total sample size
  average_total_sample_size <- lapply(total_sample_size, mean)


  ## Efficiency
  efficiency <- lapply(design, function(dd) c(efficiency=g * fptp[[dd]]['power1'] /
average_total_sample_size[[dd]]))
```

```
  inverse_efficiency <- lapply(design, function(dd) 1/efficiency[[dd]])


  ## Coverage and bias
  coveragebiasmean <-lapply(coveragebias,rowMeans,na.rm=TRUE)
  cbmse <-lapply(coveragebias,function(cb) apply(cb, 1, function(cbtmp)
    mean((cbtmp-mean(cbtmp,na.rm = TRUE))^2,na.rm = TRUE)))


  output <- list(test=fptp, mean_samplesize=average_total_sample_size,
efficiency=efficiency,cbmse=cbmse,
                 mean_coveragebias=coveragebiasmean,coveragebias=coveragebias)


  return(output)
}
```