Using Confidence Distributions in Final and Interim Analyses for Single-Arm Studies or Platform Trials Consisting of Single-Arm Studies

Günter Heimann[1] | **Peter Jacko**[2,3] | **Tom Parke**[2]

[1]Advanced Quantitative Sciences / Pharmacometrics, Novartis Pharma AG, Basel, Switzerland

[2]Berry Consultants, Abingdon, United Kingdom

[3]Department of Management Science, Lancaster University, Lancaster, United Kingdom

Correspondence
Corresponding author Günter Heimann,
Email: guenter.heimann@gmx.de

Present address
Novartis Pharma AG, Fabrikstrasse 2, 4002 Basel, Switzerland

Abstract

Confidence distributions are a frequentist alternative to the Bayesian posterior distribution. These confidence distributions have received more attention in the recent past because of their simplicity. In rare diseases, oncology, or in pediatric drug development, single-arm trials or platform trials consisting of a series of single-arm trials are increasingly being used, both to establish proof-of-concept and to provide pivotal evidence for a marketing application. Often, these single-arm trials are designed as two-stage designs, or they include sequential or continuous monitoring approaches. They are analyzed using standard frequentist, Bayesian, or other methods. In this paper, we describe how to define analysis strategies based on confidence distributions for such single-arm trials or for platform trials that consist of a series of single arm trials. We focus on binary endpoints and show how to define the corresponding decision rules for final and interim analyses, and how to derive their operating characteristics exactly, e.g., without simulation. Our approach uses predictive probabilities rather than conditional probabilities (as with stochastic curtailment) to define the interim decision rules. It can be applied to platform, basket, and umbrella trials that consist of a series of single arm trials, but also to stand-alone single arm trials.

KEYWORDS
platform trial, master protocols, single-arm study, confidence distribution, interim analysis

1 | INTRODUCTION

Master protocol trials such as umbrella, platform and basket trials, are becoming more popular for rare diseases[1,2,3] because they offer an increase in efficiency that may outweigh the methodological, logistical, operational, and regulatory issues associated with them[4,5]. Meyer et al.[3] conducted a comprehensive literature review and found that 29 out of the 50 master protocols included in their review (covering the period from 1999 to 2019) were designed as a collection of single-arm studies, and that 26 out of these 50 trials were using a binary endpoint.

In rare diseases or in pediatric drug development, single-arm trials are increasingly being used[6,7,8] as pivotal evidence in an application for marketing authorization. The European Medicines Agency (EMA) published a reflection paper[9] on using single-arm trials as pivotal evidence for this purpose, stating that "the considerations in this reflection paper extend to trials that contain more than one arm, but do not randomise to a control for a formal comparison". They continue to say that "an example for such a trial would be a particular kind of platform trial where several investigational treatment arms are included but which are not formally compared, and which can be viewed as a series of single-arm trials".

Often, these single-arm trials are designed as two-stage designs or include sequential or continuous monitoring[10,11] to allow early stopping for futility and/or efficacy. Frequentist approaches to analyze such trials include standard hypothesis tests for binary endpoints, stochastic curtailment approaches[11], and others. Bayesian approaches[12] use the posterior distribution to define a dual success criterion that combines evidence of superiority against a historical control rate, and evidence that the response rate exceeds a certain threshold[13,14,15].

The EU Patient-Centric Clinical Trial Platform (EU-PEARL) project[16,17], which was a partnership between the public and private sectors under the umbrella of the Innovative Medicines Initiative (IMI) of the EU Commission, proposed a framework for the future conduct of platform trials and developed master protocols in four diseases, including neurofibromatosis. The two master protocols that have been developed for neurofibromatosis type 1 (NF1) and type 2 (NF2)[18] as part of the EU-PEARL project can be regarded as examples for platform-basket trial which are a series of single-arm trials, with a binary endpoint, with the option to use on the the single-arm trials for pivotal evidence, and with the option to include an interim analysis.

Confidence distributions are a frequentist alternative to Bayesian posterior distributions. Although confidence distributions are a relatively old concept and were already mentioned by Cox in 1958[19], they are not well known and have not been used much until recently. Xie and Singh[20] provided a precise definition of confidence distributions along with an overview of potential applications. We also refer readers to a paper by Marschner[21] for a more recent review, and to the book by Schweder and Hjort[22] for a more comprehensive and technical introduction.

Confidence distributions summarize the knowledge and uncertainty about an unknown model parameter in the form of a probability distribution on the parameter space, just like a posterior distribution, without assuming that the parameter of interest is a random variable. They are often simple to derive either from pivots or from a bootstrap distribution, are closely linked to p-value functions, and one can obtain confidence intervals for any level from them[20]. Confidence distributions do not require a prior distribution, yet they allow one to define similar success criteria as with posterior distributions.

Our work was motivated by the two platform-basket trials in neurofibramatosis (NF) that have been developed as part of the EU-PEARL project[16,17]. These two platform trials can be seen as a collection of single-arm proof-of-concept or single-arm phase I or phase II studies[18], and their primary endpoint is binary. For each of these single-arm studies, an interim analysis may be included depending on the sponsor's request. Unlike the two review papers by Xie and Singh[20] and Marschner[21] we present an application of confidence distributions to a specific situation.

In this paper, we provide statistical analysis strategies based on confidence distributions for single-arm proof-of-concept (PoC) or single-arm phase I or phase II studies, and for master protocol trials that are a series of single-arm studies. We focus on a binary endpoint here, but extensions of the concept to continuous or other endpoints are possible. We also focus on interim analyses rules which allow for early stopping because of projected lack of success at the final analysis, but similar rules for declaring futility or success at interim analyses based on confidence distributions could be defined as well. Obviously, these would require adjustment of the type I error in case of early stopping for success.

This paper is organized as follows. We provide more details about the motivating example of platform-basket trials in NF in Section 2, and introduce the corresponding decision rules and success criteria in Section 3. We explain how to derive the operating characteristics for these decision rules without simulation in Section 4. Section 5 concludes the main part of our paper. All technical statistical derivations are deferred to the supplementary material (Section A). The software for calculations and visualizations used in this paper can be found at GitHub[23], which includes plain R code, link to a shiny app, and an MS Excel spreadsheet.

2 | MOTIVATING EXAMPLE: TWO PLATFORM-BASKET TRIALS FOR NEUROFIBROMATOSIS

Neurofibromatosis type 1 (NF1), neurofibromatosis type 2 (NF2) related Schwannomatosis, and non-NF2 related Schwannomatosis (SWN) are rare genetic disorders with an increased risk of developing nerve sheath tumors. These tumors are mostly benign, but can transform into malignant tumors. There are many manifestations of these three subtypes of neurofibromatosis, which vary considerably, and require different ways to measure response[24,25,26]. Inclusion and exclusion criteria differ between the manifestations, and possibly also between interventions (such as age restrictions, concomitant medications that should be excluded, etc.).

In such rare diseases with high variability between manifestations and different subtypes, it can be time consuming and difficult to perform separate clinical trials[27], and master protocol trials may be an alternative. As part of the EU-PEARL partnership[16], two platform-basket trials were designed, one covering four manifestations of NF1, and the other covering five manifestations of NF2 and SWN. Both the NF1 and the NF2/SWN platform trials are designed to include multiple interventions, and these may be added at different points in time by different sponsors. Some interventions may be tested against several manifestations of NF1 or NF2/SWN, and others may only be tested against one manifestation. Comparisons between interventions are not planned.

The primary endpoint for all manifestations is always a binary endpoint (response/non-response). However, the definition of this binary endpoint differs between manifestations (but not across interventions when tested in the same manifestation). For many of the manifestations, the endpoint is defined based on tumor growth data obtained from MRI, applying criteria as REiNS[28,29], RAPNO[30], or criteria similar to the RECIST 1.1 criteria[31]. In some manifestations, improvement of pain or visual acuity is part of the responder definition. For some manifestations, the endpoint is observed as early as after 3 months, for others only after 12 months.

Therefore, the platform-basket trial may be regarded as a collection of single-arm studies, with each combination of an intervention and a manifestation being a separate subtrial. Each subtrial is analyzed separately. The primary objective of each subtrial is to demonstrate proof-of-concept (PoC), e.g., that the intervention shows activity against the manifestation in which it is being tested. There is also an option to run such a subtrial with a more confirmative objective in mind. Since this would usually require a larger sample size, the EU-PEARL platform-basket trial design also allows for interim analyses to stop a subtrial early for projected lack of success, to avoid treating patients with a potentially futile intervention.

For a more detailed description and justification of the platform-basket trial design, we refer to the paper by Dhaenens et al.[18]. The rationale for selecting the manifestations for the NF1 trial has been described in Dhaenens et al.[26]. The two master protocols are published online[16], together with a detailed technical report on the analysis and the corresponding operating characteristics[32].

3 | DECISION RULES FOR A SINGLE-ARM STUDY BASED ON CONFIDENCE DISTRIBUTIONS

In this section, we introduce the decision rules for a single-arm proof-of-concept study and for a single-arm study that is intended to provide pivotal evidence for registration (registrational objective). These decision rules correspond to the dual-criterion design introduced by others[14,15] for proof-of-concept and phase II studies. They combine statistical significance with clinical relevance, and "have (been) applied in many phase II designs"[15]. We refer to Roychoudhury et al. 2020[15] for a more extensive introduction and discussion of these designs.

The primary endpoint is assumed to be a binary endpoint, as discussed in Section 2. Different to Roychoudhury et al. 2020[15], the decision rules in this paper will be based on a confidence distribution for the unknown response rate, rather than on a Bayesian posterior distribution.

3.1 | Decision Rules for the Final Analysis

In the case of a single-arm subtrial with a binary endpoint, a confidence distribution is a data-driven distribution on the interval extending from 0% response (ineffective intervention) to 100% response. It is often described by its confidence density (see Figure 1), which is centered around the observed response rate. The width of the confidence density is a function of the observed variance and the sample size. We use an asymptotic approximation based on an asymptotic pivot[20] here.

In order to define these decision rules precisely, we need some notation. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ be a column vector of N independent and identically distributed binary random variables, where $Y_\nu = 1$ ($Y_\nu = 0$) indicates response (non-response) with regard to the primary endpoint for the ν-th subject within a subtrial. The Y_ν follow Bernoulli distributions with response probability p, and their sum $Y_\cdot = \sum_{\nu=1}^{N} Y_\nu$ follows a binomial $\mathtt{Bin}(N, p)$ distribution with parameters N and p.

A confidence distribution for the unknown model parameter $p \in [0, 1]$ is a data-driven distribution on the corresponding parameter space $[0, 1]$. To be precise, a function $H = H(p \mid \boldsymbol{Y})$ of the data and the parameter p is a confidence distribution if the following two conditions apply:

1. $H(\cdot \mid \boldsymbol{Y})$ is a continuous distribution function on $[0, 1]$ for each possible outcome $\boldsymbol{Y}$.
2. The random variable $H(p \mid \boldsymbol{Y})$ is uniformly distributed on $[0, 1]$ if p is the true underlying response rate.

If either one or both of these conditions holds asymptotically, then one speaks of an asymptotic confidence distribution.

The function $H(\cdot \mid \boldsymbol{Y})$ is the distribution function of the confidence distribution. We often use the corresponding density (called *confidence density*) to represent the confidence distribution, as shown in Figure 1.

A confidence distribution summarizes the knowledge on the unknown parameter obtained from the data. Confidence intervals, estimators, or p-values for hypothesis tests can be derived from such a confidence distribution[20].

In the model considered here, we can derive an asymptotic confidence distribution

$$H_N(p \mid Y_\cdot) := \Phi\left(\frac{p - \widehat{p}_N(Y_\cdot)}{\widehat{\sigma}_N(Y_\cdot)}\right) \tag{1}$$

from the asymptotic pivot $\frac{\widehat{p}_N(Y_\cdot)-p}{\widehat{\sigma}_N(Y_\cdot)}$. This is a common approach to constructing confidence distributions[20]. The function Φ in (1) refers to the distribution function of a standard normal distribution, so that $H_N(\cdot \mid Y_\cdot)$ corresponds to a normal distribution with mean equal to the observed response rate

$$\widehat{p}_N = \widehat{p}_N(Y_\cdot) = \overline{Y}_\cdot = \frac{1}{N}\sum_{\nu=1}^{N} Y_\nu \ , \tag{2}$$

and with variance

$$\widehat{\sigma}_N^2 = \widehat{\sigma}_N^2(Y_\cdot) = \max\left\{\frac{\overline{Y}_\cdot\left(1-\overline{Y}_\cdot\right)}{N}, \frac{\frac{1}{4N^2}\left(1-\frac{1}{4N^2}\right)}{N}\right\} \ . \tag{3}$$

Note that the variance defined in (3) is always positive, even when $\overline{Y}_\cdot = 0$ or $\overline{Y}_\cdot = 1$. It equals the usual variance estimator for $\widehat{p}_N$ for all other possible values of $\overline{Y}_\cdot$.

$H_N(p \mid Y_\cdot)$ as defined in (1) is an asymptotic confidence distribution, because both of the conditions needed to define a confidence distribution apply asymptotically. For example, the central limit theorem implies that $\frac{\widehat{p}_N(Y_\cdot)-p}{\widehat{\sigma}_N(Y_\cdot)}$ is asymptotically distributed as a standard normal distribution if p is the true parameter. Therefore, $H_N(p \mid Y_\cdot)$ is asymptotically distributed as a uniform distribution on $[0, 1]$ if p is the true parameter, so that the second condition applies asymptotically.

With this, we can now define the decision rule for the final analysis, which corresponds to the dual-criterion design[15] For each subtrial, a *desired response rate* p_1, and an *ineffective response rate* response rate p_0 need to be set, with $p_0 \leq p_1$. Proof-of-concept or success is declared during the final analysis if

$$H_N(p_0 \mid Y_\cdot) < \alpha \quad \text{and} \quad H_N(p_1 \mid Y_\cdot) < \beta \tag{4}$$

is true for some pre-specified α and β satisfying $0 < \alpha < 0.5 \leq 1 - \beta < 1$. Similarly, one can define a decision rule to declare futility at the final analysis if

$$H_N(p_0 \mid Y_\cdot) > \gamma \tag{5}$$

is true for some pre-specified γ satisfying $0.5 \leq \gamma < 1$.

The first condition in (4) corresponds to a standard hypothesis test for $H_0 : p \leq p_0$ versus $H_1 : p > p_0$ with significance level equal to α. It is fulfilled if and only if H_0 is rejected using the asymptotic test statistic $\frac{\widehat{p}_N(Y_\cdot)-p_0}{\widehat{\sigma}_N(Y_\cdot)}$. The second condition in (4) corresponds to a condition on clinical relevance via the observed response rate. For example, with $\beta = 0.5$, the second condition is equivalent to the observed response rate $\widehat{p}_N(Y_\cdot)$ being strictly greater than the desired response rate p_1.

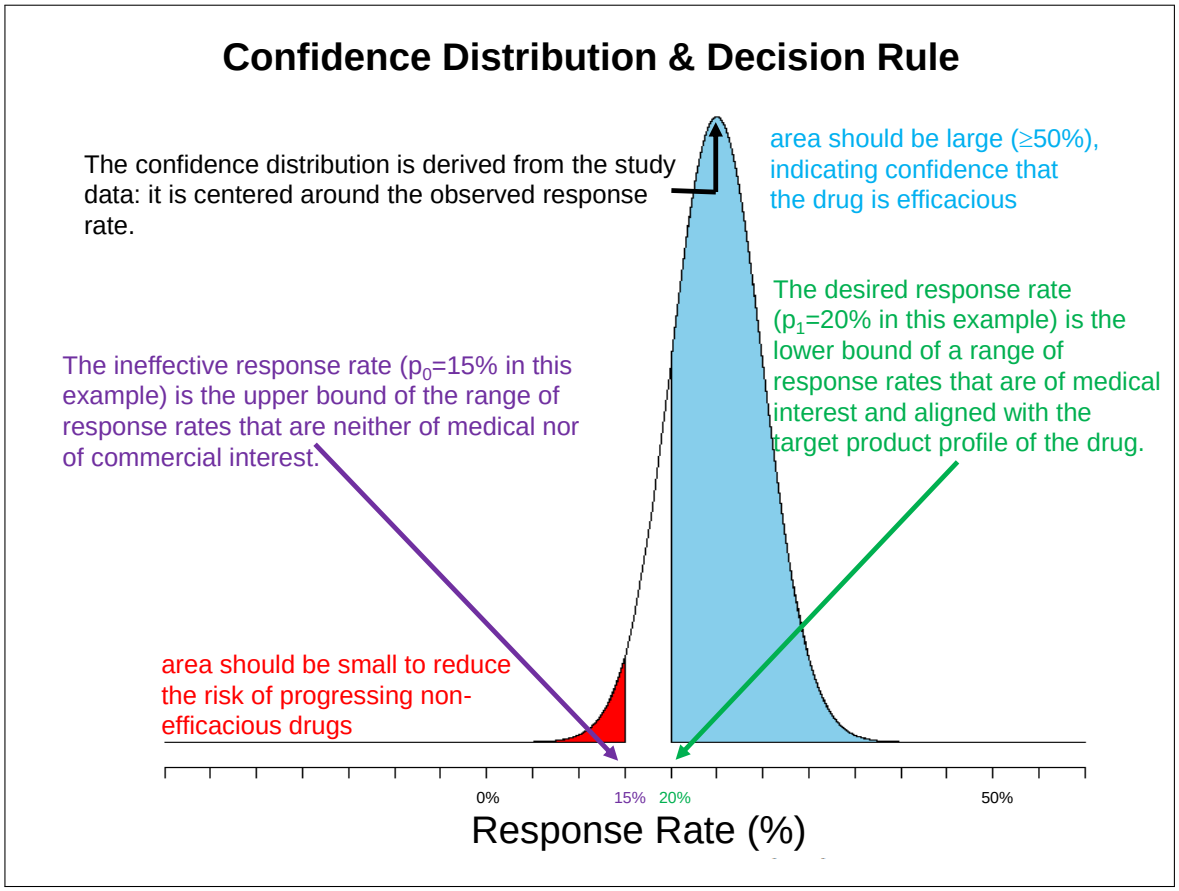
These decision rules are illustrated graphically in Figure 1. The first condition in (4) corresponds to the red area under the confidence density being smaller than α, and the second condition in (4) corresponds to the blue area being larger than $1 - \beta$. The condition in (5) corresponds to the red area under the confidence density being larger than γ.

The red area corresponds to the probability that the confidence distribution assigns to the range of ineffective response rates (ranging from 0 to $p_0 = 15\%$ in the example presented in Figure 1), and the blue area under the confidence density curve corresponds to the probability that the confidence distribution assigns to the range of response rates above the desired response rate (ranging from $p_1 = 20\%$ to 100% in Figure 1). In the two platform-basket trials described in Section 2, the choice of these parameters may vary across the different subtrials.

Reasonable choices for α, β and γ could be $\alpha = 0.2$, $\beta = 0.5$, and $\gamma = 0.5$ in a subtrial with the objective to demonstrate proof-of-concept. For a subtrial with a registrational objective, smaller values such as $\alpha = 0.05$ or $\alpha = 0.025$ may be a better choice, because the decision rule (4) corresponds to a hypothesis test at significance level α for $H_0 : p \leq p_0$ versus $H_1 : p > p_0$ with the additional condition that the observed response rate needs to exceed p_1.

In a recent paper[33] on the treatment of BRAF-aberrant or neurofibromatosis type 1-associated low-grade glioma, response rates below 0.1 were regarded as ineffective, and response rates above 0.3 were regarded as effective. The paper reported the outcome of a cohort of $N = 25$ patients with NF1-related low-grade glioma, with $Y_\cdot = 10$ of these patients being responders. The decision rule (4) declares proof-of-concept or success when applied to these data, with $p_0 = 0.1$, $p_1 = 0.3$, $\alpha = 0.1$ (the significance level that Fangusaro et al.[33] used in their analysis), and $\beta = 0.5$.

FIGURE 1 Graphical representation of the decision rule.



We prove in Section A.1 of the supplementary material that the decision rule (4) is equivalent to $Y_\cdot \geq c_N$, with c_N being defined to be the smallest integer such that

$$\widehat{p}_N(y) > \max\left\{ p_0 - \widehat{\sigma}_N(y)\Phi^{-1}(\alpha)\,,\; p_1 - \widehat{\sigma}_N(y)\Phi^{-1}(\beta) \right\} \tag{6}$$

is true for all $y \geq c_N$. Note that $c_N = c_N(p_0, p_1, \alpha, \beta)$ is a function of the parameters p_0, p_1, α, and β.

Similarly, the decision rule (5) is equivalent to $Y_\cdot \leq d_N$, with $d_N = d_N(p_0, p_1, \gamma)$ being defined as the largest integer such that

$$\widehat{p}_N(y) < p_0 - \widehat{\sigma}_N(y)\Phi^{-1}(\gamma) \quad \textit{for all} \quad y \leq d_N\,. \tag{7}$$

This is also proven in Section A.1.

3.2 | Decision Rules for Interim Analyses

Many different types of interim decision rules can be defined. These different options include stopping rules to declare success or failure at the interim analysis as well as stopping rules to decide whether to continue the enrollment to the subtrial, or whether to stop it (without declaring success or failure at the interim). Here we describe the latter case, which is part of the proposed analysis for the EU-PEARL master protocol[32].

If the primary objective of the single-arm subtrial is to declare proof-of-concept at the final analysis, the sample size may be very small, and an interim analysis may not make sense, whatever the criterion. In such a case we suggest monitoring the number of responders, and to declare proof-of-concept as soon as the observed number of responders equals the number of responders

c_N (as defined in (6)) that is needed to declare proof-of-concept at the final analysis. Whether or not to stop the subtrial at this point in time depends upon the context: if there are other subtrials with other promising interventions ongoing as part of the master protocol trial, early stopping of a successful subtrial may be useful to save resources and enable learning about the other interventions. However, if no other subtrial is ongoing for the same manifestation, early stopping of a successful intervention may not be useful and prevent patients from having an active treatment option.

If the primary objective of the single-arm subtrial is to use the data for regulatory purposes, for example to get the marketing authorization for an additional indication in a rare disease for the intervention of interest, the overall sample size N of the subtrial will still be larger. In this case, we propose to include an interim analysis after $n < N$ subjects, and to stop the subtrial if the likelihood to be successful at the final analysis, given the data at the interim analysis, is too small. This approach would prevent the sponsor from running a larger trial in a rare disease with no or little chance to be successful, and it would also allow to allocate future patients to more promising interventions. Since there is no intention to declare success early and to stop the trial after the interim analysis, there is no need to adjust the type I error here.

In order to precisely define such a decision rule, and to quantify the "likelihood to be successful at the final analysis given the data at the interim analysis", we need more notation. Let $Y_{[1\cdot]} = \sum_{\nu=1}^{n} Y_\nu$ be the sum of the first n observations, which is distributed according to a binomial distribution $\texttt{Bin}(n, p)$ with parameters n and p, and $Y_{[2\cdot]} = \sum_{\nu=n+1}^{N} Y_\nu$ be the sum of the second $N-n$ observations, distributed according to a binomial distribution $\texttt{Bin}(N-n, p)$ with parameters $N-n$ and p. Both of these distributions depend on the same unknown parameter p. At the interim analysis, the unknown parameter p can be estimated by $\widehat{p}_n(Y_{[1\cdot]}) = \frac{Y_{[1\cdot]}}{n}$. We also use $\widehat{p}_n(y) = \frac{y}{n}$ for $0 \le y \le n$, $b(y \mid n, p) := \binom{n}{y} p^y (1-p)^{n-y}$ to denote the binomial probabilities, and $B(\cdot \mid n, p)$ to denote the cumulative distribution function of a binomial distribution with parameters p and n.

With this, we can define a distribution

$$\texttt{Pred}_{N,n,Y_{[1\cdot]}} := \sum_{y=0}^{n} b(y \mid n, \widehat{p}_n(Y_{[1\cdot]}) \texttt{Bin}(N-n, \widehat{p}_n(y)) \,, \tag{8}$$

which is a mixture of binomial distributions with parameters $N-n$ and $\widehat{p}_n(y)$, and with weights $b(y \mid n, \widehat{p}_n(Y_{[1\cdot]})$ from a $\texttt{Bin}(n, \widehat{p}_n(Y_{[1\cdot]})$ distribution. The corresponding density function is

$$g(z \mid N, n, Y_{[1\cdot]}) = \sum_{y=0}^{n} b(y \mid n, \widehat{p}(Y_{[1\cdot]})\, b(z \mid N-n, \widehat{p}(y)) \tag{9}$$

for $z \in \{0, \ldots, N-n\}$, and we will use $G(\cdot \mid N, n, Y_{[1\cdot]})$ to denote the corresponding cumulative distribution function. This distribution will be used to predict the number of responders $Y_{[2\cdot]}$ among the second $N-n$ observations.

Given that $Y_{[2\cdot]}$ is a binomial $\texttt{Bin}(N-n, p)$ random variable, we could have used the binomial $\texttt{Bin}(N-n, \widehat{p}_n(Y_{[1\cdot]})$ with estimated parameter to predict the number of responders $Y_{[2\cdot]}$. In Section A.2 of the supplementary material, we discuss that the mixture distribution (8) is a better choice for this purpose, because it usually provides a better approximation to the true $\texttt{Bin}(N-n, p)$ as compared to $\texttt{Bin}(N-n, \widehat{p}_n(Y_{[1\cdot]})$.

We now use the mixture distribution (8) to quantify the *likelihood to be successful during the final analysis given the data at the interim analysis*. The subtrial will continue after the interim analysis if this likelihood exceeds a predefined threshold δ, e.g., if

$$\mathbb{P}rob_{N,n,Y_{[1\cdot]}}\{1 \le y \le N-n \,:\, H_N(p_0 \mid y + Y_{[1\cdot]}) < \alpha \text{ and } H_N(p_1 \mid y + Y_{[1\cdot]}) < \beta\} \;>\; \delta \,, \tag{10}$$

for some $0 < \delta < 1$, and it will be stopped if the condition above is not met. This decision rule can be expressed equivalently as

$$1 - G(c_N - Y_{[1\cdot]} - 1 \mid N, n, Y_{[1\cdot]}) > \delta \,, \tag{11}$$

see Section A.4 of the supplementary material. Since stopping does not occur to declare success early, no multiplicity adjustment of the significance level is needed.

Our approach to define interim decision criteria is conceptually similar to stochastic curtailment [34]. In stochastic curtailment, a conditional distribution with pre-specified parameter p_0 is used to evaluate the likelihood to meet a criterion at the end of the trial. In the situation with a binary endpoint, this conditional distribution is a binomial distribution. In Section 4.2 we show that stochastic curtailment with a fixed parameter can be less efficient than using the predictive distribution. Even when using the conditional distribution with an estimated parameter value, a mixture distribution like the predictive distribution $G(\cdot \mid N, n, Y_{[1\cdot]})$ is usually a better choice to evaluate the probability in (10) or (11), as shown in Section A.2 of the supplementary material.

Similar approaches have been used and discussed by various authors[35,36,37,38]. Unlike our approach, these authors use the Bayesian posterior predictive distribution instead of a frequentist predictive distribution. These approaches lead to similar results as ours when using a non-informative prior, as discussed in Section A.3.

We would like to point out that stopping the trial at the interim analysis because condition (10) (or equivalently (11)) is not met does not mean to declare futility at the interim analysis. If one wanted to stop early to declare futility, one would need to use a condition like (5) at the interim analysis. Using conditions (10) or (11) simply means to stop early because the data available render it unlikely to declare success at the end. The corresponding decision to stop or to continue depends upon the quality of the intervention (i.e., the true response rate) and on the decision criteria (i.e., the p_0, p_1, α, and β). For example, consider two subtrials using the same intervention and the same sample size for final and interim analysis. Assume further that both subtrials use the same decision criteria, except that the parameter p_1 for the first subtrial is smaller than the corresponding parameter for the second subtrial. In this case, the predictive probability to be successful at the end will tend to be larger in the first subtrial. One could even find more extreme examples, where the intervention in the second subtrial has a better true response probability as compared to the intervention in the first subtrial, but where the predictive probability to declare success at the end is larger for the first subtrial. This could happen if there is a much stricter success criterion required for the second subtrial.

4 | OPERATING CHARACTERISTICS FOR THE DECISION RULES

We now examine the operating characteristics of the decision rules defined in Section 3 for both the final and the interim analyses. We do this for sample sizes that correspond to a single-arm proof-of-concept study or a single-arm phase I or phase II study, and we show how sample size calculations can be done. The operating characteristics are all calculated exactly, without simulation. The results in this Section immediately extend to master-protocol studies that are a series of single-arm trials.

4.1 | Operating Characteristics for the Final Analysis

In this section we present the operating characteristics for the decision rules defined in (4) and (5) assuming that there is no interim analysis that could lead to stopping the subtrial early. The operating characteristics represent the probability that the conditions in (4) and (5) will be met as a function of the true response rate of the intervention.

Figure 2 displays the operating characteristics for these decision rules for a sample size of $N = 25$, an ineffective response rate $p_0 = 0.1$, a desired response rate $p_1 = 0.3$, $\alpha = 0.05$, and $\beta = \gamma = 0.5$.

The solid green curve corresponds to the decision rule (4) and shows the probability of declaring proof-of-concept after $N = 25$ subjects have completed the trial as a function of the unknown response rate of the intervention. This function equals

$$p \longrightarrow \sum_{y=c_N}^{N} b(y \mid N, p) \;=\; 1 - B(c_N - 1 \mid N, p)\,, \tag{12}$$

with c_N defined as in (6), see Section A.1 of the supplementary material.

The blue curve corresponds to the decision rule (5) and shows the probability of declaring futility after $N = 25$ subjects as a function of the response rate. This function equals

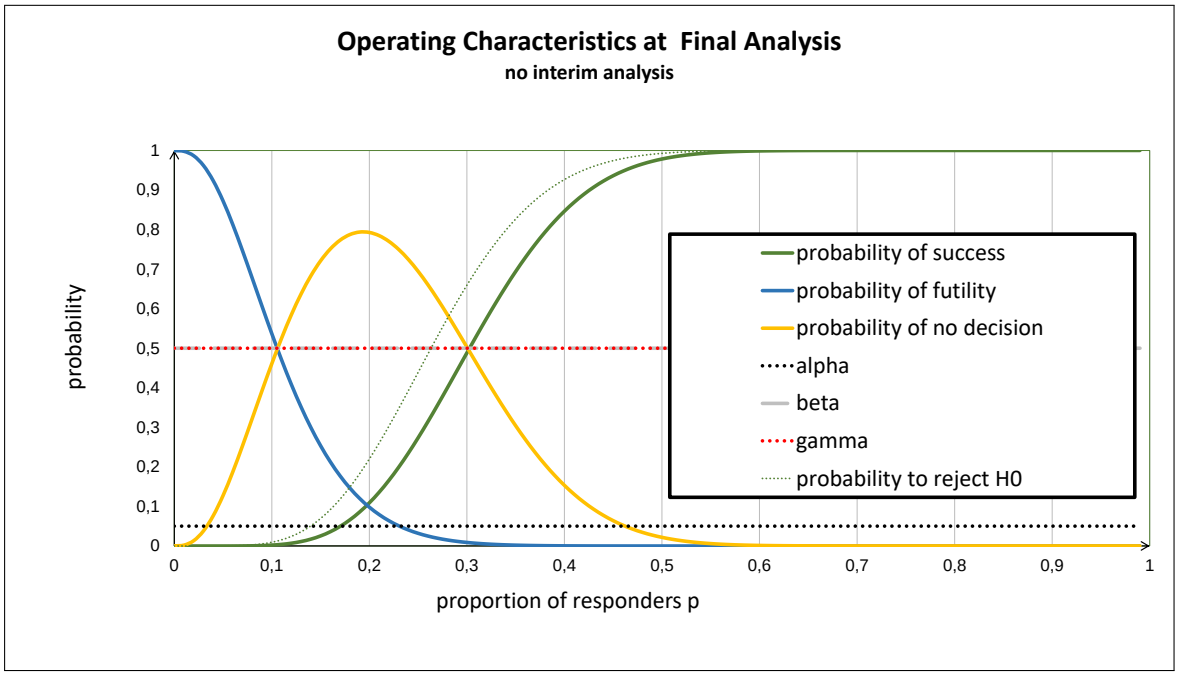
$$p \longrightarrow \sum_{y=0}^{d_N} b(y \mid N, p) \;=\; B(d_N \mid N, p)\,, \tag{13}$$

with d_N defined as in (11), see Section A.1 of the supplementary material. As one would expect, this probability is large for ineffective interventions, and very small for desired ones.

The yellow curve shows the probability of no decision (i.e., one can neither declare proof-of-concept nor futility) as a function of the response rate. It is equal to one minus the other two probabilities. The distinction between declaring futility and no decision is that in the former case, there is sufficient evidence that the intervention does not work, whereas in the latter case there is not enough information to declare either futility or success.

The dotted green curve ("probability to reject H_0") is the power function of a hypothesis test for $H_0 : p_0 \leq \alpha$ versus $H_1 : p_0 > \alpha$, and serves as a reference here. The hypothesis test corresponds to the first of the two conditions in decision rule (4).

FIGURE 2 Operating characteristics for $p_0 = 0.1$, $p_1 = 0.3$, $N = 25$, $\alpha = 0.05$, and $\beta = \gamma = 0.5$.



Therefore, this curve is always larger than or equal to the solid green curve, and their difference quantifies the loss in power caused by the second condition in (4).

We show in Figure A1 that the operating characteristics of a Bayesian decision rule that replaces the confidence distribution H_N in (4) by a corresponding Bayesian posterior distribution is identical or very similar to that in Figure 2, when using a non-informative prior. Details are described in Section A.3 of the supplementary material.

In the scenario displayed in Figure 2, the probability of declaring proof-of-concept is close to zero for an ineffective intervention with underlying response rate smaller than $p_0 = 0.1$. However, for efficacious interventions with response rates larger than $p_1 = 0.3$, the probability to declare proof-of-concept is larger than 50%, and larger than 80% if the intervention has a response rate equal to or greater than 0.4. Note that the observed response rates in the trial were above 0.4 for both strata that were reported by Fangusaro et al.[33]

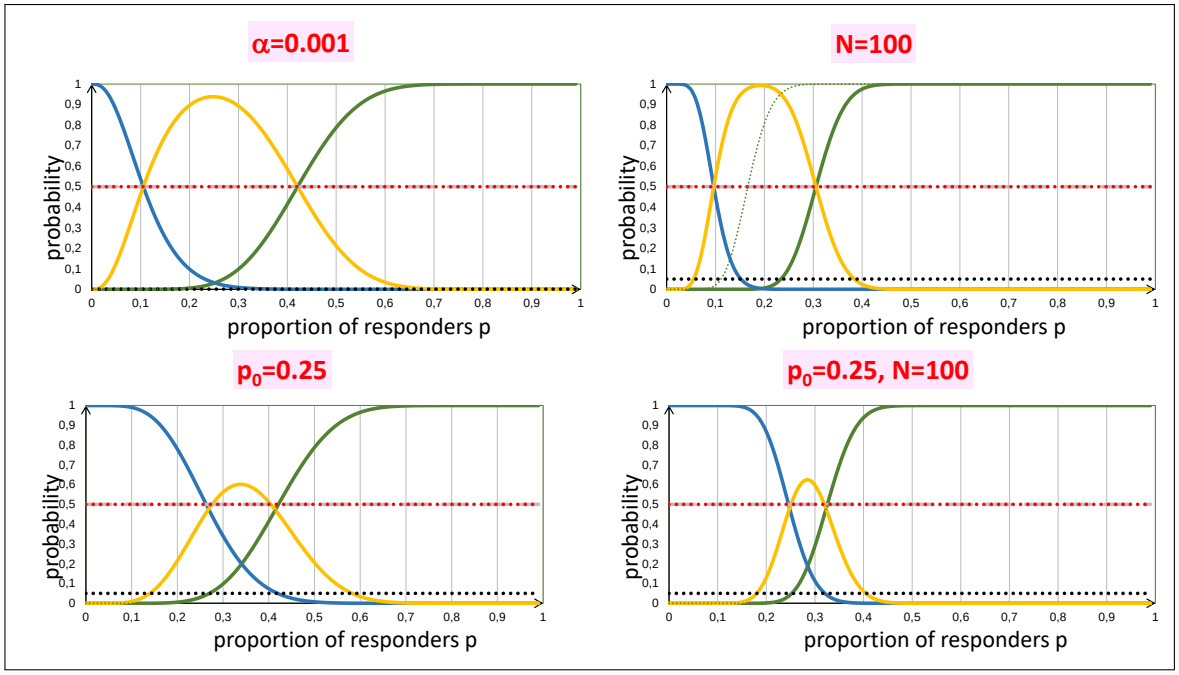
By construction of the decision rule, the probability of declaring proof-of-concept cannot exceed $1 - \beta$, if the response rate of the intervention was exactly at the boundary of the efficacious space. Similarly, the probability of declaring futility if the response rate is exactly at the boundary of the ineffective space cannot exceed γ. In the example in Figure 2, the corresponding probabilities almost achieve these limits.

Note that the probability of no decision is quite high for response rates near the boundary. For example, the probability of no decision exceeds 30% for an intervention with response rate $p = 0.35$. This can be improved by increasing the sample size. In the upper right panel of Figure 3 we show the operating characteristics of the decision rule with $N = 100$ (while all other parameters are as in Figure 2). Now the curves are much steeper, and the probability of no decision has decreased to approximately 17% for an intervention with response rate $p = 0.35$.

The two left panels in Figure 3 show that the probability of declaring proof-of-concept decreases considerably if one changes α or p_0. In both cases, the probability of declaring proof-of-concept based on $N = 25$ subjects has decreased from above 80% to below 50% even if the intervention has a response rate equal to $p = 0.4$, if we decrease α (from 0.05 to 0.001) or increase p_0 (from 0.1 to 0.25), keeping all other parameters the same as in Figure 2. If such scenarios were of interest, an increase in sample size would be needed. The bottom right panel shows that the probability of declaring proof-of-concept is again well above 80%, if $p_0 = 0.25$, when the sample size is increased to $N = 100$.

Note that the dotted and the solid green curve coincide in three of the four panels of Figure 3. This is because the second condition in (4) has no impact on the decision in these scenarios, and hence does not cause a power loss. This is the case for very

FIGURE 3 Operating characteristics for the same scenario as Figure 2 ($p_0 = 0.1$, $p_1 = 0.3$, $N = 25$, $\alpha = 0.05$, and $\beta = \gamma = 0.5$) with small variations indicated in the respective panels in red.



small values of α or when the difference $p_1 - p_0$ between the desired and the ineffective response rates is very small. On the other hand, the power loss is considerable for large sample sizes, as shown in the upper right panel of this figure. The second condition prevents one from overpowering a study and from having a significant result when the intervention's response rate is actually small. The second condition ensures that one can only declare success when a clinical relevant effect (an effect larger than p_1) is observed in the study. The probability to declare success is approximately 0.5 if the true response rate is equal to p_1, regardless of sample size.

Figure 4 displays the probability of declaring proof-of-concept (blue curve).

$$N \longrightarrow \sum_{y=c_N}^{N} b(y \mid N, p_*) \;=\; 1 - B(c_N - 1 \mid N, p_*) \tag{14}$$

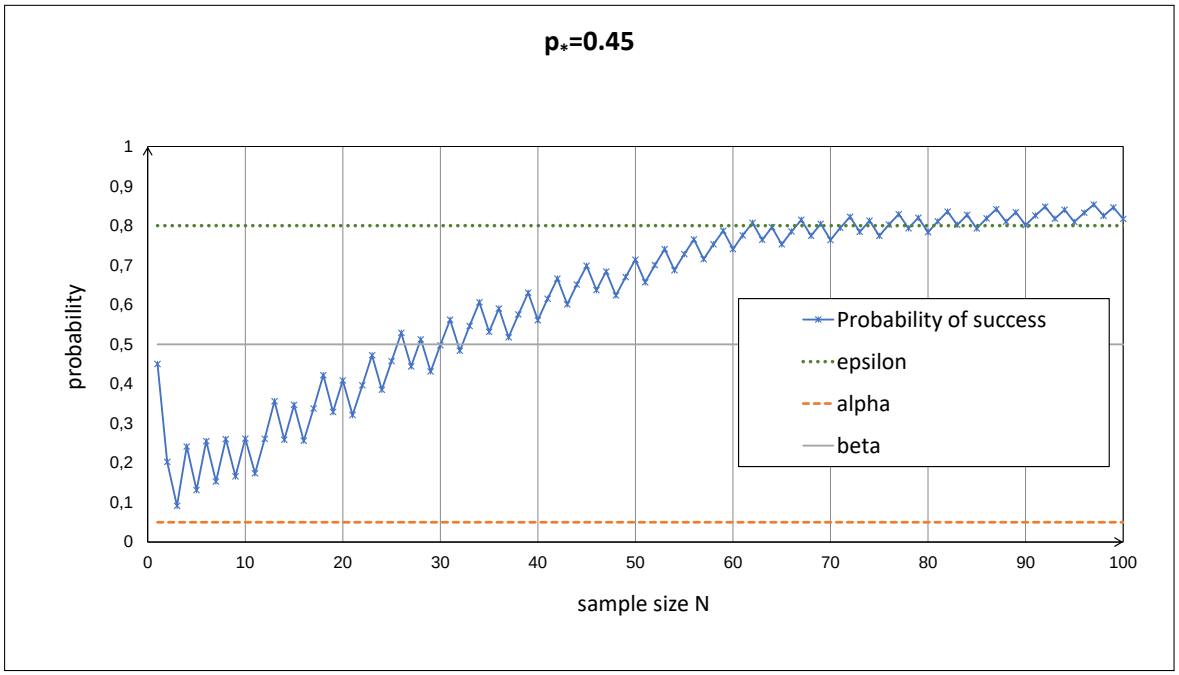
as a function of sample size N for a selected parameter value $p_* > p_1$, see Section A.1 of the supplementary material. Linear interpolation is applied between the discrete points $(N, 1 - B(c_N - 1 \mid N, p_*))$ for better readability. This function is not monotonically increasing, due to the discrete nature of the binary response endpoint. This type of figure or function can be used to find a sample size N_* which meets the condition $\sum_{y=c_{N_*}}^{N_*} b(y \mid N_*, p_*) \geq \epsilon$.

In Figure 4, we show a sample size calculation for a phase II study with a hypothetical new intervention. Given that Fangusaro et al.[33] reported a response rate larger than 0.4 for selumetinib, we use a stricter definition for an efficacious intervention, and we selected $p_1 = 0.4$ (instead of 0.3) as the desired response rate. Similarly, we selected $p_0 = 0.3$ (instead of 0.1) as the upper limit of the range of ineffective response rates. The parameters $\alpha = 0.05$ and $\beta = \gamma = 0.5$ were as before. The probability of declaring proof-of-concept at the final analysis is calculated in Figure 4 assuming that the true response rate of the hypothetical new intervention is $p_* = 0.45$.

With this choice we want to have a chance of $\epsilon = 80\%$ or more to declare proof of concept at the final analysis. One could consider N_*, the smallest sample size that achieves this, or N^*, the smallest sample size that achieves this for all $N \geq N^*$. From Figure 4 one can see that $N_* = 62$ and $N^* = 86$.

The power as a function of sample size is non-monotonic in sample size. This phenomenon has already been described elsewhere[39]. Because of this non-monotonicity, some care should be taken when selecting the appropriate sample size for a

FIGURE 4 Probability of success as a function of sample size for $p_0 = 0.3$, $p_1 = 0.4$, $p_* = 0.45$, $\alpha = 0.05$, $\beta = 0.5$, and $\epsilon = 0.8$.



subtrial. One should not blindly select N_* or N^*, but one needs to carefully assess whether there are better choices that may only require a small increase or even allow for a small decrease in sample size. However, one also needs to keep in mind that missing values may occur, and that the sample size at the end of the subtrial may not be the same as what was planned.

4.2 | Operating Characteristics in the Case of Interim Analyses

We start with the operating characteristics for a phase I or phase II study with an interim analysis to stop early because of projected lack of success after $n < N$ patients. The decision decision rule that has been defined in (10) is used.

In Figure 5 we present the operating characteristics of this interim decision rule for the scenario discussed at the end of the previous section. The total sample size for the study is $N = 62$, and the interim sample size is $n = 25$. The ineffective response rate is $p_0 = 0.3$, and the desired response rate is $p_1 = 0.4$. The parameters $\alpha = 0.05$, $\beta = 0.5$, and $\delta = 0.5$ were used to obtain these operating characteristics. These numbers were selected as a reflection of the data reported by Fangusaro et al.[33], see also the second last paragraph of the previous section.

The blue curve shows the probability of continuing after $n = 25$ subjects have been observed as a function of the true underlying response rate of the investigational intervention. It can be calculated according to

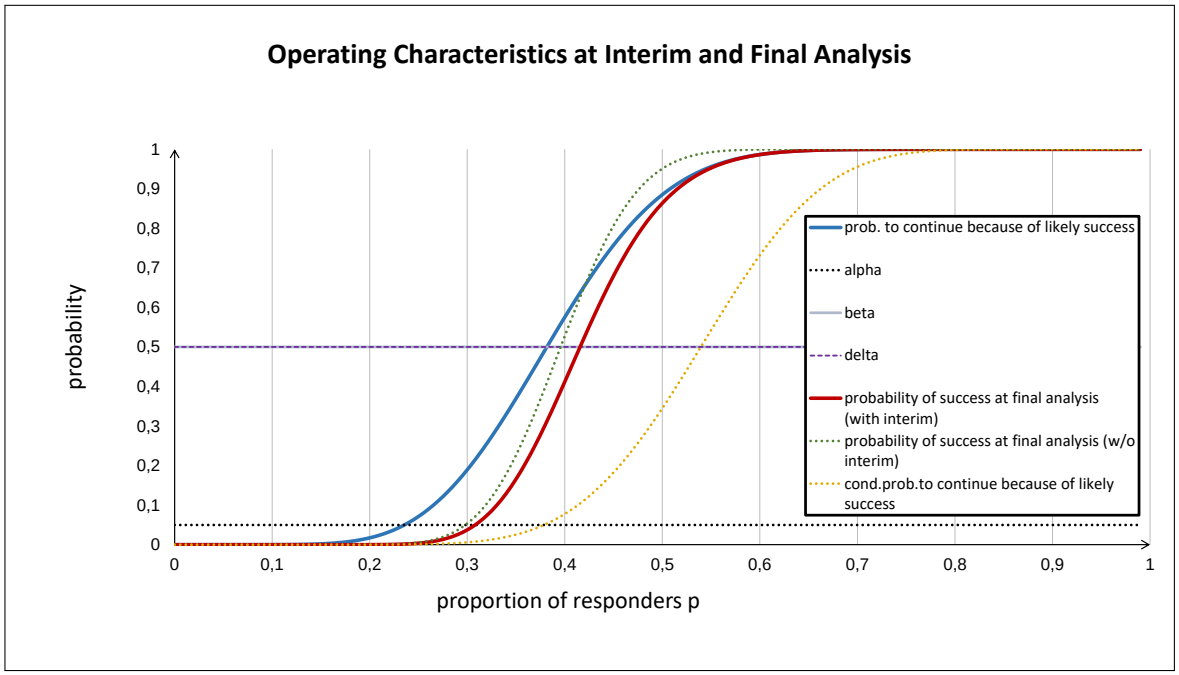
$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n, p) \mathbb{1}\{y \mid 1 - G(c_N - y - 1 \mid N, n, y) > \delta\} \tag{15}$$

because of the equivalence of (10) and (11), see Section A.4 of the supplementary material. The indicator function $\mathbb{1}\{A\}$ equals 1 if the condition A is met, and 0 otherwise.

The dotted yellow curve in Figure 5 shows the probability of continuing when using stochastic curtailment to define analogous rules to (10) or (11). We show in Section A.6 of the supplementary material how to define this curve, and we will demonstrate that the yellow curve equals

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n, p) \mathbb{1}\{y \mid 1 - B(c_N - y - 1 \mid N - n, p_0) > \delta\} \ . \tag{16}$$

FIGURE 5 Operating characteristics for $p_0 = 0.3$, $p_1 = 0.4$, $N = 62$, $n = 25$, $\alpha = 0.05$, $\beta = 0.5$, and $\delta = 0.5$.



The red curve shows the probability of declaring success during the final analysis with $N = 62$ subjects in the subtrial when there was an interim analysis after $n = 25$ subjects. It can be calculated according to

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n, p)(1 - B(c_N - y - 1 \mid N - n, p))\mathbb{1}\{y \mid 1 - G(c_N - y - 1 \mid N, n, y) > \delta\}\,, \tag{17}$$

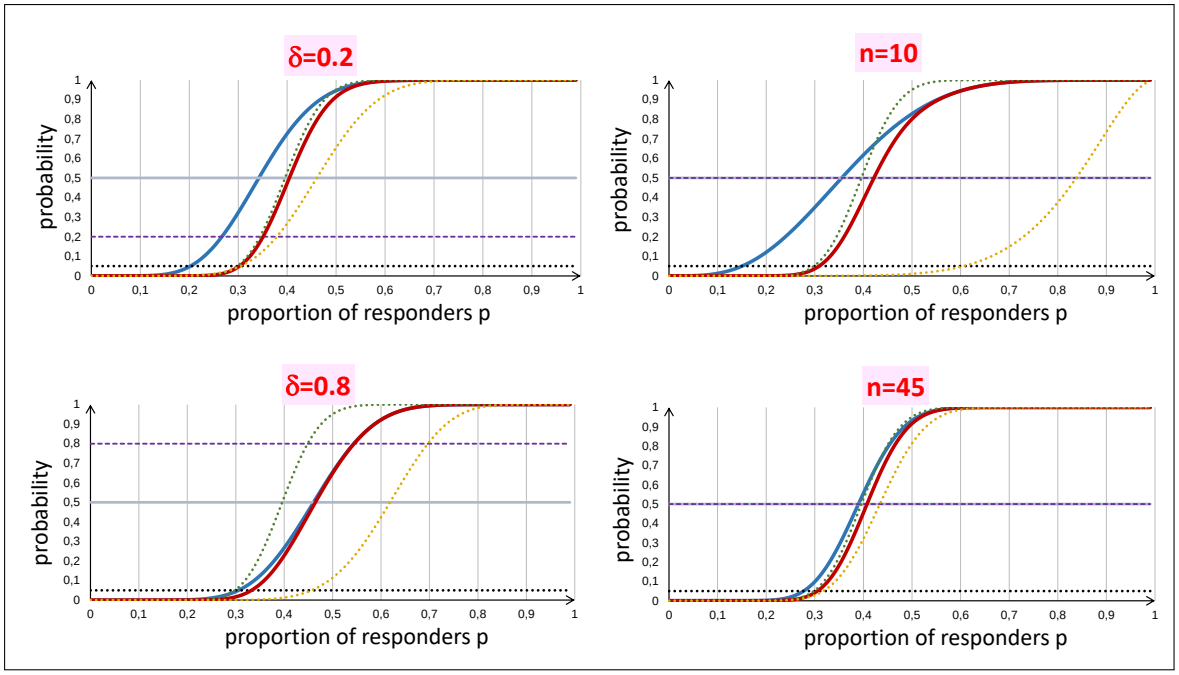
see Section A.5 of the supplementary material. The dotted green curve shows the probability of declaring success during the final analysis based on $N = 62$ subjects, when no interim analysis was done. This curve serves as a reference to quantify the loss of power imposed by an interim analysis.

In the scenario presented in Figure 5, the loss of power introduced by the interim analysis (e.g., the difference between the dotted green and the red curve) is moderate. One can also see that the approach based on stochastic curtailment is much more conservative than the predictive probability approach. This is illustrated by the difference between the blue and the dotted yellow curve.

In Figure 5 we have used $\delta = 0.5$. This means that one continues the subtrial after $n = 25$ patients, if the chance of success at the end is predicted to be larger than 50%. At a first glance, this seems to be a very low threshold to overcome, and one may want to increase δ. However, this will lead to a considerable loss in power during the final analysis, as demonstrated in the bottom left panel of Figure 6 where $\delta = 0.8$ and all other parameters are as in Figure 5. Decreasing δ further may also lead to undesirable results. The top left panel in Figure 6 shows the situation for $\delta = 0.2$. The power loss is now almost gone, but the probability of continuing after the interim analysis is very high, even for low response rates (e.g., for the ineffective response rate p_0 the probability to continue is almost 0.4). This contradicts the purpose of the interim analysis, which should stop further recruitment if an intervention is ineffective. When selecting $\delta = 0.5$ as in Figure 5, there is a good balance between power loss and probability of continuing (which is 0.2 for p_0).

The right panels of Figure 6 show what happens when the interim sample size is varied. If the interim sample size is reduced to $n = 10$, with all other parameters equal to Figure 5, the power loss increases, and the probability of continuing for low response rates increases. Hence, one should not perform an interim analysis too early. For larger interim sample sizes (as $n = 45$ in the bottom right panel), the power loss is almost gone, and the probability of stopping at the interim analysis is very close to the probability of declaring success at the end of the subtrial. This is not surprising, as more than two thirds of the subjects would

FIGURE 6 Operating characteristics for the same scenario as Figure 5 ($p_0 = 0.3$, $p_1 = 0.4$, $N = 62$, $n = 25$, $\alpha = 0.05$, $\beta = 0.5$, and $\delta = 0.5$) with small variations indicated in the respective panels in red.



have been observed in this scenario. However, for ineffective interventions one wants to stop the subtrial early, and waiting with an interim until two thirds of the subjects have been observed is counter-intuitive.

One can also see that the stochastic curtailment approach performs better if the sample size for the interim is large, or if δ is small. However, its performance is terrible for small interim sample sizes and for large values of δ.

5 | DISCUSSION

In this paper we presented statistical analysis strategies for single-arm Proof-of-Concept (PoC) or single-arm phase I or phase II studies with a binary endpoint. These analysis strategies can also be used in master protocols that are a series of single-arm subtrials. Our work was motivated by the EU-PEARL master protocols, which can be understood as a series of single-arm subtrials. Similar approaches can be defined for randomized controlled trials with a binary endpoint, or to corresponding master protocols that include a control.

The analysis strategies presented here are based on confidence distributions, a frequentist analogue of Bayesian posterior distributions. Using a distribution to summarize the data from a clinical study allows one to define decision criteria which are used in Bayesian statistics, such as (4) or (5). Such criteria are often being used for proof-of-concept studies in the pharmaceutical industry [15]. In such proof-of-concept studies, hypothesis tests that control the type I error at the usual $\alpha = 0.05$ or $\alpha = 0.025$ level are usually infeasible due to the large sample sizes required. More generally, "scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold"[40]. Specifically in proof-of-concept or phase II studies, decision makers would like to see a certain treatment effect before moving to the next phase, so that a decision rule like (4) is often convenient [14,15].

Often, Bayesian posterior distributions are used to define such decision rules, and these require prior distributions to be specified. However, there is usually little prior information when designing a proof-of-concept study, so that non-informative priors would have to be used. Confidence distributions do not require the experimenter to define a prior distribution, and their operating characteristics are identical or very similar to those of the Bayesian procedures when using non-informative priors. Moreover, their operating characteristics can often be calculated directly, as in the binary case, so that one does not

need simulation to obtain them. Therefore, confidence distributions represent a simple alternative to using Bayesian posterior distributions.

Uncontrolled single-arm studies are used in many rare diseases and in pediatric drug development, either because there is no adequate control treatment, or because a placebo controlled study is considered not ethical. The master protocol that motivated this work is such an example. The different manifestations of neurofibromatosis are rare diseases with childhood onset, and without an established standard of care. A standard development program is not feasible in such situations.

For phase I or phase II studies, interim analyses that allow one to stop a study early for futility or projected lack of success at the final analysis are a useful tool to prevent patients from being treated with an ineffective intervention, and to save costs. These questions are particularly important in the context of the platform trials that motivated this work, because patients are rare and stopping one subtrial early may allow us to test a more promising intervention more quickly.

In this paper we have discussed how to combine decision rules that are based on confidence distributions with predictive distributions to define interim decision rules. We have shown that these rules have satisfactory performance, and that the power loss can be small as compared to a situation when no interim analysis is performed. This depends on the settings, and one needs to carefully tune the parameters of the decision rule when planning such studies. We have also shown that using predictive distributions is a more efficient approach than using stochastic curtailment. The predictive probability approach can also be used to define adaptive randomization schemes in platform trials. We will discuss this in more detail in a separate paper.

The type of interim decision we consider here is different from what is usually done in interim analyses, where early stopping is either for futility or for success (the latter requiring type I error adjustment). Here we suggest to stop for lack of projected success, which is different from a decision to stop for futility. Stopping for futility means that the interim data indicate that the intervention does not work, while stopping for lack of projected success means that the study is unlikely to reach its objective (even though the interim data could indicate that the intervention works very well).

Given that the overall sample size N for a subtrial (even when the objective is to provide pivotal evidence) is likely to be rather small in a rare disease, even when the objective is to use the data as pivotal evidence, a corresponding interim sample size n would be even smaller. A more traditional interim analysis based on a decision rule like (4) and designed to declare success early will not have a lot of power in such a situation.

We have focused on decision rules that were motivated by the platform trials in neurofibromatosis. But our approach can be expanded into many different types of interim decision rules, including stopping early to declare success, or to declare futility. Our approach can also be used for sequential decision making.

One can also use the predictive probabilities defined in (10) to define adaptive randomization rules if there is more than one investigational intervention (i.e., more than one single-arm subtrial) open for enrollment in parallel for a given manifestation. For example, if there are S subtrials recruiting, with n_s subjects being available in subtrial s, one can randomly allocate the subject $n_1 + \cdots + n_S + 1$ to one of these subtrials using an allocation probability which is informed by the predictive probabilities $\pi_s(n_s)$ (with $\pi_s(n_s)$ being the predictive probability for the s-th subtrial after n_s subjects as defined in (10)). These ideas will be further discussed in a separate paper[41].

All the ideas in this paper focused on master protocols that are a series of single-arm subtrials, or on single-arm trials, and on a binary endpoint. This focus was implied by the EU-PEARL master protocols, which were our starting point. However, the approach can easily be extended to trials with a control, or to master protocols which include a control when there is a binary endpoint. Conceptually, the approach can also be applied to other types of endpoints (continuous, time-to-event, etc.) because confidence distributions can always be defined[20]. Depending on complexity, exact calculations may not be possible any more, and the operating characteristics may have to be determined via simulation.

FINANCIAL DISCLOSURE

This research was originated and partly supported by the EU-PEARL project that received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 853966-2; this Joint Undertaking received support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

CONFLICT OF INTEREST

Peter Jacko and Tom Parke are employees of Berry Consultants, a consulting company that specializes in the design, conduct, oversight, and analysis of adaptive and platform clinical trials. Günter Heimann was an employee of Novartis Pharma AG when this work was mostly done and owned stocks of this company.

REFERENCES

1. Woodcock J, LaVange L. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *New England Journal of Medicine.* 2017;377:62-70. doi: 10.1056/NEJMra1510062
2. Park J, Siden E, Zoratti M, et al. Systematic Review of Basket Trials, Umbrella Trials, and Platform trials: A Landscape Analysis of Master Protocols. *Trials.* 2019;20. doi: 10.1186/s13063-019-3664-1
3. Meyer EL, Mesenbrink P, Dunger-Baldauf C, et al. The Evolution of Master Protocol Clinical Trial Designs: A Systematic Literature Review. *Clinical Therapeutics.* 2025;42(7):1330-1360. doi: 10.1016/j.clinthera.2020.05.010
4. Kaizer A, Koopmeiners J, Chen N, Hobbs B. Statistical Design Considerations for Trials that Study Multiple Indications. *Statistical Methods in Medical Research.* 2020;30:096228022097518. doi: 10.1177/0962280220975187
5. Li X, Lu C, Broglia C, et al. Current Usage and Challenges of Master Protocols—Based on Survey Results by ASA BIOP Oncology Methodology Working Group Master Protocol Sub-team. *Annals of Translational Medicine.* 2020;10(18). doi: 10.21037/atm-21-6139
6. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ Open.* 2016;6(6). doi: 10.1136/bmjopen-2016-011666
7. Zou D, Zhang E, Wu S, Rose V. Use of Single-Arm Trials in FDA Approvals of Treatments in Relapsed or Refractory B-Cell Lymphoma. *Blood.* 2023;142(Supplement 1):7250-7250. doi: 10.1016/j.cct.2023.107200
8. Wang M, Ma H, Shi Y, Ni H, Qin C, Ji C. Single-arm clinical trials: design, ethics, principles. *BMJ Supportive & Palliative Care.* 2024. doi: 10.1136/spcare-2024-004984
9. European Medicines Agency *Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation application*. 2024.
10. Simon RM. Optimal Two-Stage Designs for Phase II Clincal Trials. *Conrolled Clinical Trials.* 1989;10:1-10.
11. Law M, Grayling MJ, Mander AP. A stochastically curtailed single-arm randomised phase II trial design for binary outcomes. *Pharmaceutical Statistics.* 2022;32(5):671-691. doi: 10.1080/105434406.2021.2009498
12. Zohar S, Teramukai S, Zhou Y. Bayesian desgn and conduct of phase II single-arm clinical trials with binary outcomes: A tutorial. *Contemporary Clinical Trials.* 2008;29(4):608-616. doi: 10.1016/j,cct.2007.11005
13. Di Scala L, Kerman J, Neuenschwander B. Collection, synthesis, and interpretation of evidence: a proof-of-concept study in COPD. *Statistics in Medicine.* 2013;32(10):1621-1634. doi: 10.1002/sim.5730
14. Fisch R, Jones I, Jones J, Kerman J, Rosenkranz G, Schmidli H. Bayesian Design of Proof-of-Concept Trials. *Therapeutic Innovation & Regulatory Science.* 2014;49:155-162. doi: 10.1177/2168479014533970
15. Roychoudhury S, Scheuer N, Neuenschwander B. Beyond p-values: A phase II dual.criterion design with statistical significance and clinical relevance. *Clinical Trials.* 2025;15(5):452-461. doi: 10.1177/740774518770661
16. EU-PEARL website . https://eu-pearl.eu/; 2020.
17. König F, Spiertz C, Millar D, et al. Current State-of-the-Art and Gaps in Platform Trials: 10 Things You Should Know, Insights from EU-PEARL. *eClinicalMedicine.* 2024;67:102384. doi: 10.1016/j.eclinm.2023.102384
18. Dhaenens B, Heimann G, Bakker A, et al. Platform Trial Design for Neurofibromatosis Type 1, NF2-related Schwannomatosis and non-NF2-related Schwannomatosis: A Potential Model for Rare Diseases. *Neuro-Oncology Practice.* 2024. doi: 10.1093/nop/npae001
19. Cox DR. Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics.* 1958;29(2):357-372.
20. Xie Mg, Singh K. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statitical Review.* 2013;81:3-39. doi: 10.1111/insr.12000
21. Marschner IC. Confidence Distributions for Treatment Effects in Clinical Trials: Posteriors Without Priors. *Statistics in Medicine.* 2024;43:1271-1289. doi: 10.1002/sim.10000
22. Schweder T, Hjort NL. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. New York, USA: Cambridge University Press, 2016.
23. Jacko P, Heimann G, Parke T. ConDistris: Designing Clinical Trials using Confidence Distributions. https://github.com/PeterJacko/ConDistris; .
24. Korf BR. Neurofibromatosis. *Handbook of Clinical Neurology.* 2013;111:333-340.
25. Blakeley JO, Plotkin SR. Therapeutic Advances for the Tumors Associated with Neurofibromatosis Type 1, Type 2, and Schwannomatosis. *Neuro-Oncology.* 2016;18(5):624-638. doi: 10.1093/neuonc/nov200
26. Dhaenens BA, Ferner RE, Bakker A, et al. Identifying Challenges in Neurofibromatosis: A Modified Delphi Procedure. *European Journal of Human Genetics.* 2021;29:1625-1633. doi: 10.1038/s41431-021-00892-z
27. Griggs R, Batshaw M, M D, Gopal-Srivastava R, Kaye E, Krischer J. Clinical Research for Rare disease: Opportunities, Challenges, and Solutions. *Molecular Genetics and Metabolism.* 2009;96(1):20-26. doi: 10.1016/j.ymgme.2008.10.003
28. Dombi E, Ardern-Holmes SL, Babovic-Vuksanovic D, et al. Recommendations for Imaging Tumor Response in Neurofibromatosis Clinical Trials. *Neurology. 2013; 81): S33-S40.*. 2013;81((21 Suppl 1):S33-S40.
29. Thalheimer R, Merker VL, Ly I, et al. Validating Techniques for Measurement of Cutaneous Neurofibromas: Recommendations for Clinical Trials. *Neurology.* 2021;97(7 Suppl 1):S32-S41. doi: 10.1212/WNL.0000000000012428
30. Fangusaro J, Witt O, Hernaíz Driever P, et al. Response Assessment in Paediatric Low-Grade Glioma: Recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) Working Group. *The Lancet Oncology.* 2020;21(6):e305-e316. doi: 10.1016/S1470-2045(20)30064-4
31. Eisenhauer E, Therasse P, Bogaerts J, et al. New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1). *European Journal of Cancer.* 2009;45:228-247. doi: 10.1016/j.ejca.2008.10.026
32. Heimann G, Jacko P, Parke T. Statistical Analysis for a Phase I Platform Trial in Neurofibromatosis. tech. rep., EU-PEARL Project; Innovative Medicines Initiative: 2023.
33. Fangusaro J, Arzu OT, Young Poussaint T, et al. Selumetinib in Paediatric Patients with BRAF-aberrant or Neurofibromatosis Type 1-Associated Recurrent, Refractory, or Progressive Low-Grade Glioma: A Multicentre, Phase 2 Trial. *Lancet Oncology.* 2019;20(7). doi: 10.1016/S1470-2045(19)30277-3
34. Lan G, Simon R, Halperin M. Stochastically Curtailed Tests in Long-Term Clinical Trials. *Communication in Statistics Part C: Sequential Analysis.* 1982;1(3):207-219. doi: 10.1080/07474948208836014

35. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring Clinical Trials: Conditional or Predictive Power. *Controlled Clinical Trials.* 1986;7:8-17. doi: 10.1016/0197-2456(86)90003-6
36. Dmitrienko A, Wang MD. Bayesian Predictive Approach to Interim Monitoring in Clinical Trials. *Statistics in Medicine.* 2006;25:2178-2195. doi: 10.1002/sim2204
37. Saville BR, Connor JT, Ayers GD, JoAnn A. The Utility of Bayesian Predictive Probabilities for Interim Monitoring of Clinical Trials. *Clinical Trials.* 2014;11(4):485-493. doi: 10.1177/1740774514531352
38. Saville BR, Berry SM. Efficiencies of Platform Clinical Trials: A Vision of the Future. *Clinical Trials.* 2016;13(3):358-366. doi: 10.1177/1740774515626362
39. Chernick MR, Y LC. The Saw-Toothed Behavior of Power Versus Sample Size and Software Solutions: Single Binomial Proportion Using Exact Methods. *The American Statistician.* 2025;56(2):149-155. doi: 10.1198/000313002317572835
40. Wasserstein RL, Lazar NA. The AS Statement on p-values: Context, Process, and Purpose. *The American Statistician.* 2016;70(2):129-133. doi: 10.1080/00031305.2016.1154108
41. Jacko P, Heimann G, Parke T. Designing Master Protocol Trials for Single-Arm Studies. *submitted to the Journal.* 2025.
42. Aitchison J. Goodness of Prediction Fit. *Biometrika.* 1975;62(3):547-554.
43. Paris RB. Incomplete Beta Functions. In: Olver FW, Lozier DM, Boisvert RF, Clark CW., eds. *NIST Handbook of Mathematical Functions*, , Cambridge University Press, 2010.

SUPPORTING INFORMATION

Additional technical derivations and proofs can be found online in the Supporting Information section at the end of this article.

How to cite this article: Heimann G., Jacko P., and Parke T.. Using Confidence Distributions in Final and Interim Analyses for Single-Arm Studies or Platform Trials Consisting of Single-Arm Studies *Statistics in Medicine* 2024;00(00):1–18.

APPENDIX

A TECHNICAL DERIVATIONS AND PROOFS

In this section we provide technical derivations and proofs for Sections 3 and 4.

A.1 Proof of (6) and (7)

The two conditions that define the decision rule (4) can be expressed equivalently as

$$H_N(p_0 \mid Y_\cdot) < \alpha \quad \Leftrightarrow \quad \widehat{p}_N(Y_\cdot) > p_0 - \widehat{\sigma}_N(Y_\cdot)\Phi^{-1}(\alpha)\,, \tag{A1}$$

and

$$H_N(p_1 \mid Y_\cdot) < \beta \quad \Leftrightarrow \quad \widehat{p}_N(Y_\cdot) > p_1 - \widehat{\sigma}_N(Y_\cdot)\Phi^{-1}(\beta)\,. \tag{A2}$$

Consider the inequality $z > p_0 - \sqrt{z(1-z)}\frac{\Phi^{-1}(\alpha)}{\sqrt{N}}$ as a function of $z \in [0,1]$. The function $z \longrightarrow z$ is continuous and increasing from 0 to 1. The function $p_0 - \sqrt{z(1-z)}\frac{\Phi^{-1}(\alpha)}{\sqrt{N}}$ is continuous and concave (because $\Phi^{-1}(\alpha)$ is negative) and equal to p_0 for $z = 0$ or $z = 1$. Therefore, the two functions must intersect, and there is exactly one value $0 < z_0 < 1$ where these functions intersect. Actually, the solution must be in the interval $\frac{1}{2N} < z_0 < 1 - \frac{1}{2N}$ unless $p_0 < \frac{1}{2N}$. With the same reasoning, there must be such a value z_1 for the second function $p_1 - \sqrt{z(1-z)}\frac{\Phi^{-1}(\beta)}{\sqrt{N}}$ (which is actually a horizontal line when $\beta = 0.5$). The corresponding inequalities in conditions (A1) and (A2) can be viewed as discrete approximations of this continuous problem, which have jumps in $\frac{1}{N}, \frac{2}{N}, \frac{3}{N}, \ldots, 1$ and are constant otherwise. The modification to bound $\widehat{\sigma}_N(Y_\cdot)$ away from zero does not disturb this argument, since this modification only applies on the intervals $[0, \frac{1}{2N}]$ and $[1 - \frac{1}{2N}, 1]$.

This shows that $c_N = c_N(p_0, p_1, \alpha, \beta)$, defined in (6) as the smallest integer such that

$$\widehat{p}_N(y) > \max\left\{\, p_0 - \widehat{\sigma}_N(y)\Phi^{-1}(\alpha)\,,\, p_1 - \widehat{\sigma}_N(y)\Phi^{-1}(\beta) \,\right\}\,, \tag{A3}$$

exists, and that (4) it is equivalent to $Y_\cdot \geq c_N$, which was to be shown. Statements (12) and (14) are a direct consequence of this.

A straightforward calculation shows that the decision rule (5) can be expressed equivalently as

$$\Phi\left(\frac{p_0 - \widehat{p}_N(Y_\cdot)}{\widehat{\sigma}_N(Y_\cdot)}\right) > \gamma \quad \Leftrightarrow \quad \widehat{p}_N(Y_\cdot) < p_0 - \widehat{\sigma}_N(Y_\cdot)\Phi^{-1}(\gamma) \ . \tag{A4}$$

One can show with the same type of reasoning as above that this rule is equivalent to $Y_\cdot \leq d_N$, with d_N being defined as the largest integer such that

$$\widehat{p}_N(y) < p_0 - \widehat{\sigma}_N(y)\Phi^{-1}(\gamma) \quad \textit{for all} \quad y \leq d_N$$

in (7). Statement (13) is a direct consequence of this.

A.2 Predictive Distributions

In the next paragraphs we will provide an explanation why the predictive distribution $G(\cdot \mid N, n, Y_{[1\cdot]})$ with density (9), i.e.,

$$g(z \mid N, n, Y_{[1\cdot]}) = \sum_{y=0}^{n} b(y \mid n, \widehat{p}(Y_{[1\cdot]})\, b(z \mid N-n, \widehat{p}(y)) \ , \tag{A5}$$

is likely to be a better choice to approximate $\texttt{Bin}(N-n, p)$ than $\texttt{Bin}(N-n, \widehat{p}_n(Y_{[1\cdot]}))$.

Assume that there are two random variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ with distributions $P_{1,\vartheta}$ and $P_{2,\vartheta}$ that depend on the same unknown parameter ϑ, and that there are corresponding densities $p_1(y_1, \vartheta)$ and $p_2(y_2, \vartheta)$. Assume further that the first random variable has been observed, and that one wants to predict the outcome of the second one, for example as part of an interim analysis. The corresponding predictive distribution should make adequate use of the observed data $\boldsymbol{Y}_1$ by estimating ϑ.

An obvious candidate for a predictive distribution for $\boldsymbol{Y}_2$ is the distribution with density

$$p_2(y_2, \widehat{\vartheta}(\boldsymbol{Y}_1)) \ , \tag{A6}$$

where the unknown parameter is replaced by an estimate $\widehat{\vartheta}(\boldsymbol{Y}_1)$ that is based on the observed data. However, it has been shown[42] that this is not a good choice, and that the distribution with density

$$\tilde{p}_2(y_2 \mid \vartheta) = \int p_2(y_2, \widehat{\vartheta}(y_1)) P_{1,\vartheta}(dy_1) \tag{A7}$$

is closer to the true distribution $P_{2,\vartheta}$ as compared to the predictive distribution (A6). In the context of prediction, closeness between two distributions P and Q is usually measured by the Kullback-Leibler divergence[42]

$$\mathcal{KL}(P, Q) = \int (\log(q(x)) - \log(p(x)))\, q(x) dx \ , \tag{A8}$$

where p and q are the densities of P and Q, respectively.

Now, the predictive distribution (A7) is of no practical use, because it depends on the unknown parameter ϑ, but it motivates

$$\tilde{p}_2(y_2 \mid \widehat{\vartheta}(\boldsymbol{Y}_1)) = \int p_2(y_2, \widehat{\vartheta}(y_1)) P_{1,\widehat{\vartheta}(\boldsymbol{Y}_1)}(dy_1) \tag{A9}$$

as an alternative to (A6). Note that in this formula, $\widehat{\vartheta}(\boldsymbol{Y}_1)$ is the observed estimate based on the first sample $\boldsymbol{Y}_1$, and $\widehat{\vartheta}(y_1)$ is a possible value of the estimate had y_1 been observed.

In the situation with binomial data $Y_{[1\cdot]} = \sum_{\nu=1}^{n} Y_\nu \sim \texttt{Bin}(n, p)$ and $Y_{[2\cdot]} = \sum_{\nu=n+1}^{N} Y_\nu \sim \texttt{Bin}(N-n, p)$, the predictive distribution (A9) corresponds to the predictive distribution $G(\cdot \mid N, n, Y_{[1\cdot]})$ with density (9), and the predictive distribution (A9) corresponds to a binomial $\texttt{Bin}(N-n, \frac{Y_{[1\cdot]}}{n})$ distribution.

A.3 Posterior Predictive Distributions

A Bayesian analogue to the decision rule (4) would be to replace the confidence distribution $H_N(p \mid Y_\cdot)$ by a posterior distribution. When using a beta prior $p \sim \texttt{Beta}(a, b)$ with parameters $a > 0$ and $b > 0$, the posterior distribution (when having observed $Y_\cdot$

responders out of N subjects) is again a beta distribution with parameters $Y_{\cdot} + a$ and $N - Y_{\cdot} + b$. We use $B_\beta(\cdot \mid Y_{\cdot} + a, N - Y_{\cdot} + b)$ for the corresponding cumulative distribution function, which is continuous.

The Bayesian posterior predictive distribution for the number of responders $Y_{[2\cdot]}$ among the second $N - n$ observations, after having observed $Y_{[1\cdot]}$ responders among the first n subjects, is a beta-binomial distribution $\texttt{BetaBin}(N-n, Y_{[1\cdot]}+a, n-Y_{[1\cdot]}+b)$ distribution with parameters $N - n$, $Y_{[1\cdot]} + a$, and $n - Y_{[1\cdot]} + b$. We use $B_{\beta b}(\cdot \mid N - n, Y_{[1\cdot]} + a, n - Y_{[1\cdot]} + b)$ to denote the corresponding cumulative distribution function, which has discrete support equal to the integers between 0 and $N - n$, and $b_{\beta b}(\cdot \mid N - n, Y_{[1\cdot]} + a, n - Y_{[1\cdot]} + b)$ to denote the corresponding density function.

With this notation, the Bayesian decision rule for the final analysis can be expressed as

$$B_\beta(p_0 \mid Y_{\cdot} + a, N - Y_{\cdot} + b) < \alpha \quad \text{and} \quad B_\beta(p_1 \mid Y_{\cdot} + a, N - Y_{\cdot} + b) < \beta \; . \tag{A10}$$

Let's use

$$B_N := \{0 \leq y \leq N \; : \; B_\beta(p_0 \mid y + a, N - y + b) < \alpha \quad \text{and} \quad B_\beta(p_1 \mid y + a, N - y + b) < \beta\} \tag{A11}$$

to denote the subset of $\{0, 1, \ldots, N\}$ that fulfills condition (A10). This subset equals $B_N = \{b_N, b_N + 1, \ldots, N\}$, where b_N is the smallest integer such that

$$B_\beta(p_0 \mid b_N + a, N - b_N + b) < \alpha \quad \text{and} \quad B_\beta(p_1 \mid b_N + a, N - b_N + b) < \beta \tag{A12}$$

holds. Such an integer b_N exists because the relationship

$$\begin{aligned} B_\beta(p \mid y + a, N - y + b) \; &= \; B_\beta(p \mid y + 1 + a, N - y - 1 + b) + \frac{p^{y+a}(1-p)^{N-y-1+b}}{(y+a)B(y+a, N-y+b)} \\ &> \; B_\beta(p \mid y + 1 + a, N - y - 1 + b) \end{aligned}$$

is true for all p[43, 8.17.19].

With this, the operating characteristic of the Bayesian decision rule for the final analysis can be expressed as

$$p \longrightarrow \sum_{y=b_N}^{N} b(y \mid N, p) = 1 - B(b_N - 1 \mid N, p) \; , \tag{A13}$$

where $b(y \mid N, p)$ is the density of a binomial distribution, and $B(y \mid N, p)$ the corresponding cumulative distribution function.

The Bayesian interim decision rule can be expressed as

$$1 - B_{\beta b}(b_N - Y_{[1\cdot]} - 1 \mid N - n, Y_{[1\cdot]} + a, n - Y_{[1\cdot]} + b) > \delta \; , \tag{A14}$$

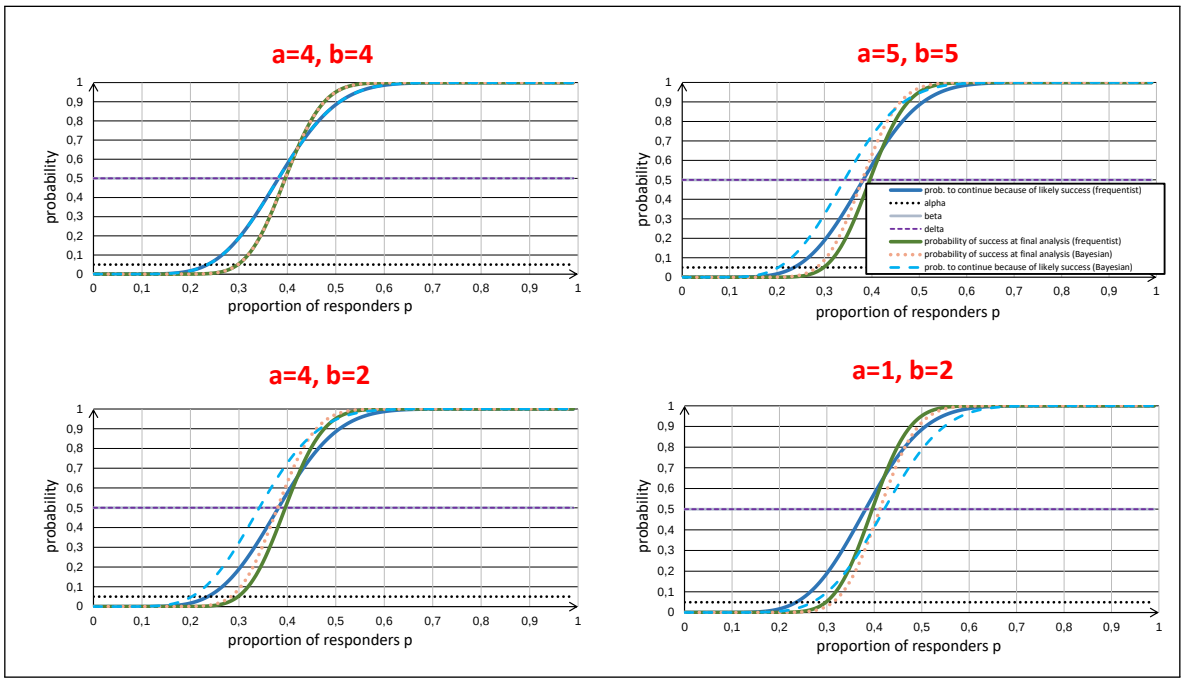
and the corresponding operating characteristic is

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n, p) \, 1\{y \; : \; 1 - B_{\beta b}(b_N - y - 1 \mid N - n, y + a, n - y + b) > \delta\} \; . \tag{A15}$$

In Figure A1 the operating characteristics of the Bayesian approach are compared with the frequentist approach. The solid green line is the probability of success at the final analysis ($N = 62$) for the frequentist procedure. The dotted yellow line is the operating characteristic of the corresponding Bayesian procedure. Similarly, the solid blue line describes the probability of continuing after the interim ($n = 25$) under the frequentist approach, and the dashed light blue line is the corresponding probability for the Bayesian procedure. The frequentist probabilities do not depend on a and b, which are the parameters of the prior distribution.

One can see in Figure A1 that for $a = 4$ and $b = 4$, there is no difference between the approaches. In fact, the curves are identical because the critical values are equal ($b_N = c_N$) in this particular example, and this is actually true for all parameters $0 < a = b \leq 4.3$, which include the well-known uninformative priors of Bayes ($a = b = 1$), Jeffrey ($a = b = 0.5$) and Haldane ($a = b \to 0$). Moreover, based on a limited computational study (not reported here) we conjecture that the curves are always identical for any choices of a and b where $\frac{a}{a+b} = p_1$, in this example, as it seems that the second condition (involving p_1 and β) is the dominating one when evaluating (A12). For other small values of a and b there are slight differences, sometimes favoring the Bayesian approach, sometimes the frequentist approach. These differences vanish when the parameters approach zero, i.e. when the prior becomes weaker, and are more pronounced when the parameters increase, i.e. when the prior becomes stronger.

FIGURE A1 Operating characteristics for frequentist (solid lines) and Bayesian (dashed or dotted lines) approach, for interim (blue and light blue lines) and final analyses (green and yellow lines), with $p_0 = 0.3$, $p_1 = 0.4$, $\alpha = 0.05$, $\beta = \delta = 0.5$, $N = 62$, $n = 25$, and different values of a and b.



A.4 Equivalence of (10) and (11)

In this section we want to prove the equivalence of (10) and (11), i.e., the equivalence of

$$\mathbb{P}rob_{N,n,Y_{[1\cdot]}}\{1 \le y \le N-n \,:\, H_N(p_0 \mid y + Y_{[1\cdot]}) < \alpha \text{ and } H_N(p_1 \mid y + Y_{[1\cdot]}) < \beta\} \;>\; \delta$$

and

$$1 - G(c_N - Y_{[1\cdot]} - 1 \mid N, n, Y_{[1\cdot]}) > \delta\,,$$

where G_N is the distribution with density (A5), see also (9).

From the definiton of c_N (see (A3) or (6)) we conclude that

$$H_N(p_0 \mid y + Y_{[1\cdot]}) < \alpha \text{ and } H_N(p_1 \mid y + Y_{[1\cdot]}) < \beta \tag{A16}$$

is equivalent to

$$y \ge c_N - Y_{[1\cdot]} \tag{A17}$$

for all integers $1 \le y \le N-n$. Since $\mathbb{P}rob_{N,n,Y_{[1\cdot]}}\{c_N - Y_{[1\cdot]}, \ldots, N-n\}$ equals $1 - \mathbb{P}rob_{N,n,Y_{[1\cdot]}}\{1, \ldots, c_N - Y_{[1\cdot]} - 1\}$ we conclude that (11) is equivalent to (10).

A.5 Proof of (15) and (17)

Given the equivalence of (10) and (11), one can continue after an interim analysis with y responders out of n subjects if $1 - G(c_N - y - 1 \mid N, n, y) > \delta$, which immediately implies

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n, p) \mathbb{1}\{y \mid 1 - G(c_N - y - 1 \mid N, n, y) > \delta\}\,,$$

i.e., (15). In order to continue after an interim analysis with y responders out of n subjects *and* to be successful at the end, one needs to observe z responders out of the following $N-n$ subjects such that $y+z \geq c_N$, see (6). The corresponding probability is $1-B(c_N-y-1 \mid N-n,p)$, which proves

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n,p)(1-B(c_N-y-1 \mid N-n,p))\mathbb{1}\{y \mid 1-G(c_N-y-1 \mid N,n,y) > \delta\} \ ,$$

i.e., (17).

A.6 Stochastic Curtailment

Stochastic curtailment was introduced by Lan et al.[34] as a conservative rule to stop early for success. Applying these ideas to our situation means to replace the predictive distribution in decision rule (10) by a conditional probability. The resulting decision rule would be

$$\mathbb{P}rob_{p_0}\{H_N(p_0 \mid Y_{[1\cdot]}+Y_{[2\cdot]}) < \alpha \text{ and } H_N(p_1 \mid Y_{[1\cdot]}+Y_{[2\cdot]}) < \beta \mid Y_{[1\cdot]}\} \ > \ \delta \ , \tag{A18}$$

which is equivalent to

$$\mathbb{P}rob_{N-n,p_0}\{1 \leq y \leq N-n \ : \ H_N(p_0 \mid y+Y_{[1\cdot]}) < \alpha \text{ and } H_N(p_1 \mid y+Y_{[1\cdot]}) < \beta\} \ > \ \delta \ . \tag{A19}$$

The probability $\mathbb{P}rob_{N-n,p_0}$ in (A19) is the binomial $\texttt{Bin}(N-n,p_0)$ probability.

Note particularly the choice of parameter p_0. The probability in (A18) and (A19) is evaluated under the assumption that the intervention is ineffective, which is a very conservative assumption when defining a decision rule to continue the trial because the results look promising.

As in Section A.4 one can now show that the condition in (A19) can be evaluated as $1-B(c_N-Y_{[1\cdot]}-1 \mid N-n,p_0(Y_{[1\cdot]}))$, which proves

$$p \longrightarrow \sum_{y=0}^{n} b(y \mid n,p)\mathbb{1}\{y \mid 1-B(c_N-y-1 \mid N-n,p_0) > \delta\} \ ,$$

i.e., (16).