Optimizing One-Sample Tests for Proportions in Single- and Two-Stage Oncology Trials

Alan David Hutson *

Posted Date: 18 July 2025

doi: 10.20944/preprints202507.1553.v1

Keywords: perturbation test; exact binomial test; small-sample power; clinical trial

 doi:10.20944/preprints202507.1553.v1

Article

Optimizing One-Sample Tests for Proportions in Single- and Two-Stage Oncology Trials

Alan D. Hutson

Roswell Park Comprehensive Cancer Center, Department of Biostatistics and Bioinformatics, Elm and Carlton Streets, Buffalo, NY 14623. email: alan.hutson@roswellpark.org

Abstract

Phase II oncology trials often rely on single-arm designs to test $H_0 : \pi = \pi_0$ versus $H_a : \pi > \pi_0$, especially when randomized trials are infeasible due to cost or disease rarity. Traditional approaches like the exact binomial test and Simon's two-stage design tend to be conservative, with actual Type I error rates falling below the nominal α due to discreteness of the underlying binomial processes. We propose a convolution-based method that combines the binomial distribution with a simulated normal variable to construct an unbiased estimator of π and ensure precise Type I error control. We derive its theoretical properties and compare its performance to exact tests in both one-stage and two-stage settings. The approach yields more efficient designs with reduced sample sizes while maintaining error rate constraints. A new two-stage design with interim futility analysis is introduced and compared to Simon's design. Real-world examples show the method's potential to reduce trial cost and duration. This work offers a flexible, efficient alternative for early-phase oncology trial design.

Keywords: perturbation test; exact binomial test; small-sample power; clinical trial

1. Introduction

Common designs for phase II oncology trials typically focus on testing the hypothesis $H_0 : \pi = \pi_0$ versus $H_a : \pi > \pi_0$ in a one-arm non-randomized setting. Although randomized trials are generally preferred, considerations such as cost, feasibility, and the rarity of certain cancer types often necessitate the use of single-arm designs. The estimated per-patient cost of conducting an oncology clinical trial was approximately $59,500, as reported by Batelle in 2013 [1]. More recent studies, particularly those involving cellular therapies, report substantially higher costs, in some cases exceeding $500,000 per treatment cycle [2,4]. Consequently, there is a critical need to optimize both phase II and phase III randomized trial designs to shorten trial duration and reduce required sample sizes. These efforts not only address escalating costs but also aim to expedite the availability of effective therapies to cancer patients. The focus of this work is towards optimizing phase II one-arm oncology trials with a binary endpoint in terms of reduced sample size and increased efficiency.

Commonly employed binary endpoints in phase II trials include objective response, complete response, and progression-free, event-free, or overall survival at fixed time points, such as 6 months or 1 year. A key feature of non-randomized, single-arm phase II trials is that, in many cancer indications, the standard-of-care population response rate is sufficiently well-characterized to serve as a comparator. If no promising signal is observed, further development, including progression to a randomized phase II or III trial, is typically not pursued.

In single-stage designs, the hypothesis about a rate or proportion, $H_0 : \pi = \pi_0$ versus $H_a : \pi > \pi_0$, is most often tested using an exact binomial test, where the Type I error rate $\leq \alpha$. In contrast, one-arm two-stage designs commonly employ Simon's two-stage design [5], using either the minimax or optimal design configurations. The minimax design minimizes the total sample size, while the optimal design minimizes the expected sample size under the null hypothesis, where with the constraints are that the Type I error rate $\leq \alpha$ and the Type II error rate $\leq \beta$. Both designs incorporate an interim futility

 doi:10.20944/preprints202507.1553.v1

analysis at a predetermined sample size to allow early termination for lack of efficacy. Historically, it is noteworthy that very similar sampling schemes were developed decades earlier in the field of quality control, referred to as double sampling plans [6], with Simon's two-stage design representing a special case of these earlier methods [7].

The issue with both single-stage and two-stage designs is that the exact Type I error rate is often considerably lower than the desired Type I error rate α and the desired power is often larger than the desired power value $1-\beta$, due to the discreteness of the underlying binomial distribution under both the null and alternative hypotheses for a given design. This phenomenon is illustrated in the so-called saw-tooth plots in Figure 1, which display the exact Type I error across a range of potential null values for π_0 with sample sizes $n = 10, 20, 30, 40$. As a result, these tests can be conservative in certain scenarios where the exact Type I error rate falls substantially below the target level α.

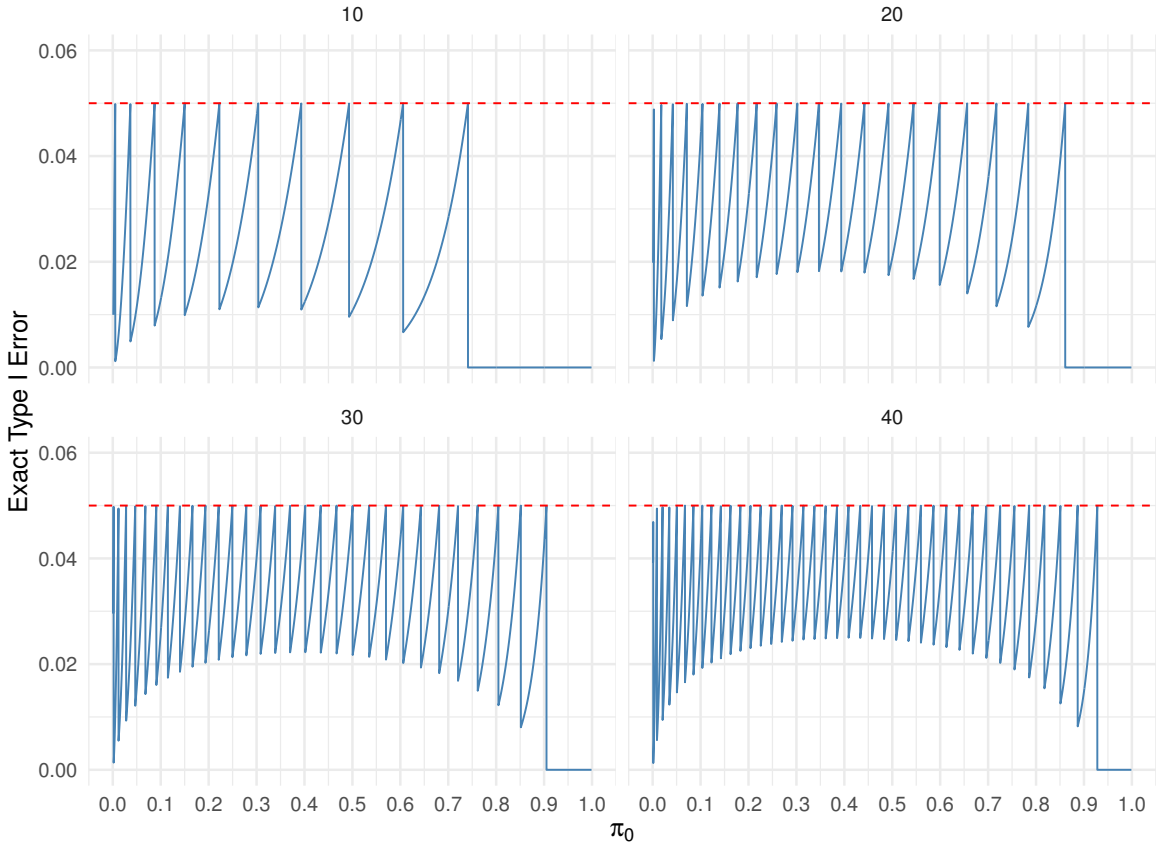

Figure 1. Type I error control for exact binomial test as a function of π_0 values, $n = 10, 20, 30, 40$.

One approach to mitigate this conservatism is to incorporate a continuity correction [9]; however, this does not eliminate the saw-tooth behavior in Type I error control. Similarly, the power function may also exhibit a saw-tooth pattern and can be non-monotonic [8]. For a fixed sample size n, there are n discrete values of π_0 corresponding to $k = 1, 2, \ldots, n$ ($k = 0$ is infeasible), at which the Type I error equals α, given by

$$\pi_0 = I_{\alpha}^{-1}(k, n-k+1),$$

where I^{-1} is the inverse regularized incomplete beta function.

Interestingly, Simon's two-stage design can exhibit the same phenomenon, where the exact Type I error rate is often much lower than the derired Type I error rate α. In this case, however, we cannot produce a saw-tooth plot as a function of π_0 alone, since Simon's two-stage design also depends on the choice of the alternative response rate, denoted $\pi_1 (> \pi_0)$. In Figure 2, we present the exact Type I error rate and power for testing $H_0 : \pi = \pi_0$ versus $H_a : \pi > \pi_0$, with $\pi_0 = 0.1$ and π_1 varying from 0.19 to 0.8, while fixing $\alpha = 0.05$ and power = 0.80 for the minimax design. As π_1 increases, the required total

 doi:10.20944/preprints202507.1553.v1

sample size decreases, thus reducing the dimensionality of the sample space for design choices. In fact, for this example, there are very few instances where the exact Type I error rate approaches the nominal α. For example, when $\pi_1 = 0.35$, the exact Type I error rate is 0.027 with a final stage sample size of $n = 18$. Increasing π_1 to 0.36 raises the exact Type I error rate to 0.044, with a corresponding final stage sample size of $n = 14$. For smaller values of n, the exact Type I error rate can drop as low as 0.015. Similarly, as the sample size decreases with increasing π_1, the exact power may deviate considerably from the target power.

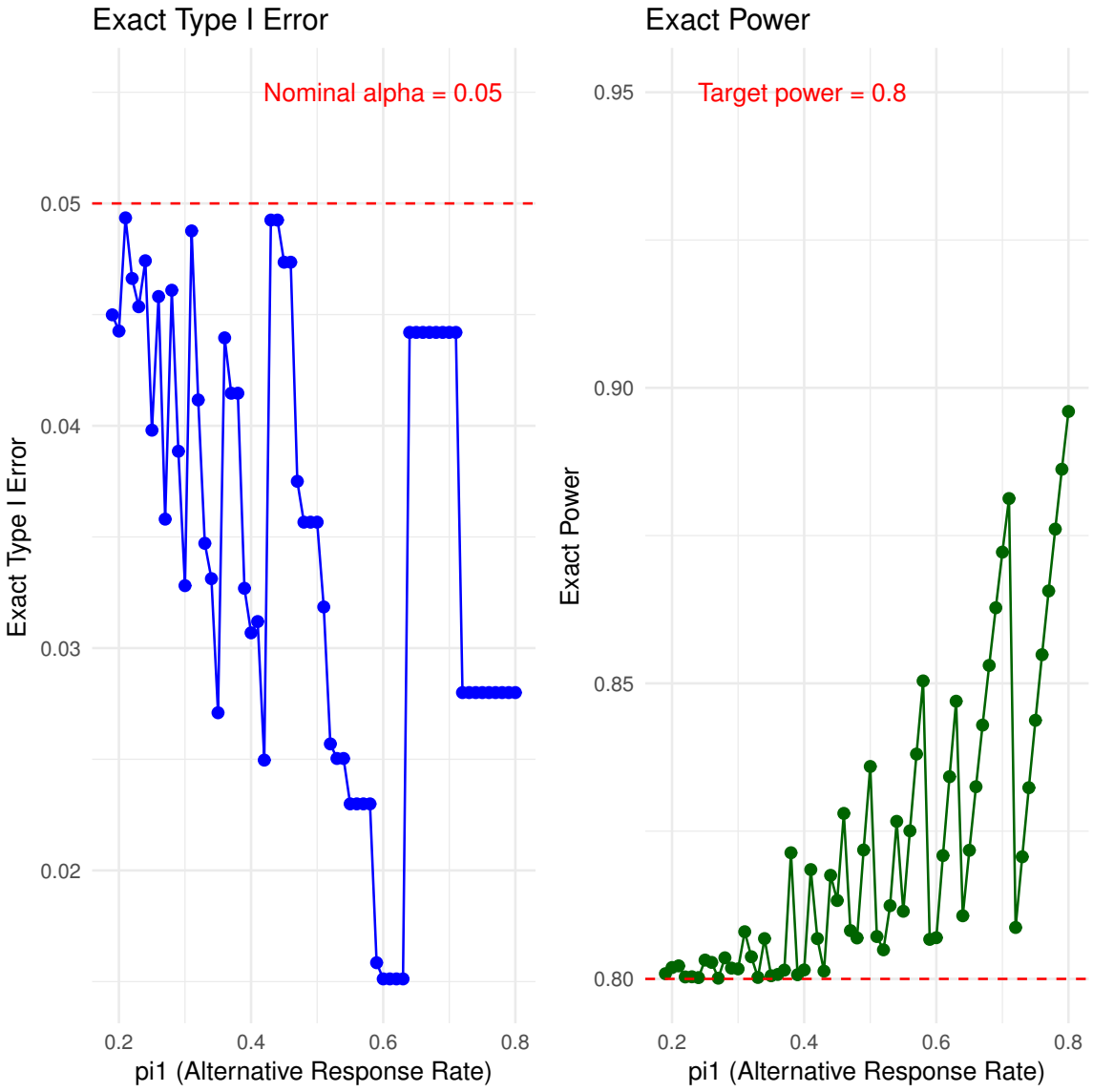

Figure 2. Type I error control and power for the mimimax Simon's two-stage design as a function of $\pi_0 = 0.1$ and $\pi_1 = 0.19$ to 0.80.

In this note, we illustrate how a straightforward convolution of a binomial random variable with a simulated normal random variable enables the construction of an unbiased estimator for the rate parameter π, and facilitates inference for $H_0 : \pi = \pi_0$ versus $H_a : \pi > \pi_0$ with precise Type I error control. This in turn can reduce sample size requirements for both one-stage and two-stage designs.

In Section 2, we define the convolution estimator, derive its density and distribution functions, outline key theoretical properties including its expectation and variance, and present a toy example to demonstrate the p-value calculation.

We then provide a detailed comparison of the new convolution-based test with the exact binomial test in terms of Type I error control and power. In Section 3, we introduce a new two-stage design with

 doi:10.20944/preprints202507.1553.v1

a futility stopping rule that also achieves precise Type I error control. A direct comparison between the convolution-based two-stage design and Simon's two-stage design is presented.

In Section 4, we provide real-world examples of both one-stage and two-stage designs, demonstrating how the convolution-based approach can reduce the cost and duration of clinical trials based on published design parameters. We conclude with final remarks.

2. Convolution Estimator

Our approach for constructing a test of the hypothesis $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, with precise Type I error control, utilizes a convolution-based method, in which synthetic continuous noise is added to discrete binomial data. Specifically, let $Y \sim \text{Binomial}(n, \pi)$ denote the binomial response over n subjects, where π is the success probability. Let $X \sim N(0, h)$ be an independent normal random variable with mean 0 and standard deviation h. We define the continuous variable $Z = Y + X$ and derive its probability density and cumulative distribution functions.

The probability density function (PDF) of Z, denoted $f_Z(z)$, is obtained by convolving the binomial probability mass function of Y with the normal density of X. Since Y is discrete and X is continuous, this results in a finite mixture of normal densities:

$$f_Z(z) = \frac{1}{h\sqrt{2\pi}} \sum_{k=0}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k} \exp\left(-\frac{(z-k)^2}{2h^2}\right), \quad z \in \mathbb{R}.$$

Each component of the mixture is centered at $k = 0, 1, \ldots, n$, with mixing weights given by the binomial probabilities $\binom{n}{k}\pi^k(1-\pi)^{n-k}$.

Similarly, the cumulative distribution function (CDF) of Z, denoted $F_Z(z) = P(Z \le z)$, is derived by conditioning on the values of Y:

$$F_Z(z) = \sum_{k=0}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k} \Phi\left(\frac{z-k}{h}\right), \quad z \in \mathbb{R}, \tag{1}$$

where $\Phi(u)$ is the standard normal CDF defined as

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

To understand the properties of the test statistic, we compare the moments of Y and Z. The expectation and variance of Y are given by

$$E[Y] = n\pi, \qquad \text{Var}(Y) = n\pi(1-\pi).$$

Since Y and X are independent, the corresponding moments of $Z = Y + X$ follow directly:

$$E[Z] = E[Y + X] = E[Y] + E[X] = n\pi + 0 = n\pi,$$

$$\text{Var}(Z) = \text{Var}(Y + X) = \text{Var}(Y) + \text{Var}(X) = n\pi(1-\pi) + h^2.$$

Thus, the mean of Z remains identical to that of Y, while the variance of Z is increased by an additive term h^2, capturing the additional variability introduced by the normal perturbation. Throughout this note, we fix the value of $h = 1/100$. Although this results in only a modest increase in the variance of Z we will demonstrate that this small adjustment is sufficient to produce tests with substantially improved power while maintaining the precise Type I error level.

To test $H_0 : \pi = \pi_0$ against $H_1 : \pi > \pi_0$ at significance level α, we reject H_0 if $Z > c$, where the critical value c is chosen to satisfy

$$1 - \sum_{k=0}^{n} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} \Phi\left(\frac{c-k}{h}\right) = \alpha. \tag{2}$$

 doi:10.20944/preprints202507.1553.v1

The solution for c is via numerical methods. Similarly, the p-value is given by

$$p = 1 - F_{Z|\pi=\pi_0}(z) = 1 - \sum_{k=0}^{n} \binom{n}{k} \pi_0^k (1-\pi_0)^{n-k} \Phi\left(\frac{z-k}{h}\right), \tag{3}$$

where z is the value of the observed convolution-based test statistic.

Under the alternative hypothesis $H_1 : \pi = \pi_1 > \pi_0$, the corresponding CDF is

$$F_{Z|\pi=\pi_1}(z) = \sum_{k=0}^{n} \binom{n}{k} \pi_1^k (1-\pi_1)^{n-k} \Phi\left(\frac{z-k}{h}\right),$$

and the power of the test at $\pi = \pi_1$ is given by

$$\text{Power}(\pi_1) = 1 - \sum_{k=0}^{n} \binom{n}{k} \pi_1^k (1-\pi_1)^{n-k} \Phi\left(\frac{c-k}{h}\right). \tag{4}$$

2.1. Toy Example

We simulated the random variables $Y \sim \text{Binomial}(n = 20, \pi = 0.3)$ and $X \sim \mathcal{N}(0, h^2)$ with $h = \frac{1}{100}$, and computed $Z = X + Y$. For each run, we calculated the convolution-based p-value and the exact binomial p-value testing $H_0 : \pi = \pi_0 \quad \text{vs} \quad H_1 : \pi > \pi_0$.

We see from Table 1 that the convolution-based p-values are consistently close to the exact binomial p-values, but tend to be slightly smaller due to the smoothing effect introduced by the normal perturbation X. The addition of this small normal noise transforms the discrete binomial distribution into a continuous mixture distribution, which leads to slightly different tail probabilities. For runs with larger observed binomial counts y, both the convolution and exact tests yield smaller p-values, indicating stronger evidence against the null hypothesis $H_0 : \pi = \pi_0$.

Table 1. Comparison of mixture-based and exact binomial p-values over three simulations.

Run	y	x	$z = x + y$	$F_Z(z)$	Convoution p-value	Exact binomial p-value
1	4	0.007968	4.007968	0.5832	0.4168	0.5886
2	11	0.014024	11.01402	0.9999	0.0001	0.0006
3	3	0.008362	3.008362	0.3701	0.6299	0.7939

2.2. Convolution Approach and Exact Binomial Test Comparison

Table 2 presents a Type I error control and power comparison between the convolution-based test and the exact binomial test for a sample size of $n = 10$, across varying null hypotheses $\pi_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and corresponding alternatives $\pi_1 \in \{\pi_0, \pi_0 + 0.1, \ldots, \min(\pi_0 + 0.4, 1)\}$. For each π_0, a critical value c was determined so that the convolution-based test controls the Type I error exactly at level $\alpha = 0.05$, as defined in equation (2). The corresponding power was computed using equation (4), noting that power equals the Type I error when $\pi_0 = \pi_1$. In contrast, the exact binomial rejection threshold k was chosen to ensure that the Type I error remained strictly below α.

Table 3 reports analogous results for a larger sample size of $n = 20$. As expected, power increases for both methods with larger n. Across both settings, the convolution-based test consistently achieves the nominal Type I error and provides greater power than the exact binomial test, particularly when the exact binomial test is overly conservative relative to Type I error control. This improved performance is attributed to the smoothing introduced by the normal perturbation, which results in a more refined and responsive rejection region. These findings underscore the practical utility of the convolution-based test in discrete-data settings, especially when measurement or process noise is present and sample sizes are limited.

 doi:10.20944/preprints202507.1553.v1

Table 2. Comparison of power between the mixture-based convolution test and the exact binomial test for sample size $n = 10$, across values of π_0 and π_1. The rejection threshold k ensures the binomial test controls Type I error strictly below $\alpha = 0.05$.

π_0	π_1	Critical c	Mixture Power	Rejection k	Binomial Power
0.1	0.1	2.9962	0.050000	4	0.012795
0.1	0.2	2.9962	0.251377	4	0.120874
0.1	0.3	2.9962	0.523352	4	0.350389
0.1	0.4	2.9962	0.757080	4	0.617719
0.1	0.5	2.9962	0.904088	4	0.828125
0.2	0.2	4.0086	0.050000	5	0.032793
0.2	0.3	4.0086	0.189362	5	0.150268
0.2	0.4	4.0086	0.415895	5	0.366897
0.2	0.5	4.0086	0.663109	5	0.623047
0.2	0.6	4.0086	0.855538	5	0.833761
0.3	0.3	5.0195	0.050000	6	0.047349
0.3	0.4	5.0195	0.171407	6	0.166239
0.3	0.5	5.0195	0.383292	6	0.376953
0.3	0.6	5.0195	0.638272	6	0.633103
0.3	0.7	5.0195	0.852383	6	0.849732
0.4	0.4	6.9878	0.050000	8	0.012295
0.4	0.5	6.9878	0.158735	8	0.054688
0.4	0.6	6.9878	0.358174	8	0.167290
0.4	0.7	6.9878	0.619691	8	0.382783
0.4	0.8	6.9878	0.856551	8	0.677800
0.5	0.5	7.9876	0.050000	9	0.010742
0.5	0.6	7.9876	0.154390	9	0.046357
0.5	0.7	7.9876	0.357879	9	0.149308
0.5	0.8	7.9876	0.645587	9	0.375810
0.5	0.9	7.9876	0.909147	9	0.736099

Table 3. Power comparison between the mixture-based convolution test and the exact binomial test for sample size $n = 20$, across values of π_0 and π_1. The rejection threshold k ensures the binomial test controls Type I error strictly below $\alpha = 0.05$.

π_0	π_1	Critical c	Mixture Power	Rejection k	Binomial Power
0.1	0.1	4.0143	0.050000	5	0.043174
0.1	0.2	4.0143	0.386941	5	0.370352
0.1	0.3	4.0143	0.772408	5	0.762492
0.1	0.4	4.0143	0.951708	5	0.949048
0.1	0.5	4.0143	0.994442	5	0.994091
0.2	0.2	7.0045	0.050000	8	0.032143
0.2	0.3	7.0045	0.281501	8	0.227728
0.2	0.4	7.0045	0.638410	8	0.584107
0.2	0.5	7.0045	0.892613	8	0.868412
0.2	0.6	7.0045	0.983738	8	0.978971
0.3	0.3	9.0186	0.050000	10	0.047962
0.3	0.4	9.0186	0.249643	10	0.244663
0.3	0.5	9.0186	0.593093	10	0.588099
0.3	0.6	9.0186	0.874692	10	0.872479
0.3	0.7	9.0186	0.983230	10	0.982855
0.4	0.4	11.9910	0.050000	13	0.021029
0.4	0.5	11.9910	0.229635	13	0.131588
0.4	0.6	11.9910	0.562559	13	0.415893
0.4	0.7	11.9910	0.865636	13	0.772272
0.4	0.8	11.9910	0.985944	13	0.967857
0.5	0.5	13.9918	0.050000	15	0.020695

 doi:10.20944/preprints202507.1553.v1

Table 3. *Cont.*

π_0	π_1	Critical c	Mixture Power	Rejection k	Binomial Power
0.5	0.6	13.9918	0.224232	15	0.125599
0.5	0.7	13.9918	0.568302	15	0.416371
0.5	0.8	13.9918	0.890702	15	0.804208
0.5	0.9	13.9918	0.995777	15	0.988747

3. Two Stage Design

Similar to the one-stage test, we construct a two-stage test of the hypothesis $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, offering precise control of the Type I error rate and allowing for early stopping due to futility. Let the total sample size be $n = n_1 + n_2$. After the first n_1 subjects have completed their endpoint assessments, a futility stopping rule is applied.

As in the one-stage design, let Z_1 denote the observed value of the convolution-based test statistic after the first n_1 subjects, and let Z_2 be the corresponding statistic based on an independent second cohort of n_2 subjects. Under the null hypothesis H_0, define the p-values as

$$p_1 = 1 - F_{Z|\pi=\pi_0}(Z_1), \quad p_2 = 1 - F_{Z|\pi=\pi_0}(Z_2),$$

where p_1 and p_2 follow a uniform $U(0,1)$ distribution by the probability integral transform. The cumulative distribution function $F_{Z|\pi=\pi_0}$ is defined in Equation(3).

Our approach uses a p-value threshold at the interim analysis and applies Stouffer's weighted z-score method for the final efficacy analysis[10]. Define the transformed statistics:

$$T_1 = \Phi^{-1}(p_1), \quad \text{based on the first } n_1 \text{ subjects,}$$

$$T_2 = \Phi^{-1}(p_2), \quad \text{based on the second } n_2 \text{ subjects,}$$

where Φ^{-1} is the quantile function of the standard normal distribution.

The combined test statistic is given by:

$$T = \frac{w_1 T_1 + w_2 T_2}{\sqrt{w_1^2 + w_2^2}},$$

where the weights are defined as $w_1 = n_1/n$ and $w_2 = n_2/n$. With this weighting scheme, the statistic T follows a standard normal distribution under H_0, i.e., $T \sim \mathcal{N}(0,1)$.

The interim futility rule is to terminate the study early if $p_1 > p_c$, where the threshold p_c may be specified by the user or selected to optimize statistical power or minimize the expected sample size,

$$\text{ESN} = n_1 p_c + n_2(1 - p_c).$$

If the study does not stop early for futility, the final p-value is computed as $p = \Phi(T)$, where Φ denotes the standard normal CDF. The null hypothesis H_0 is rejected if $p < \alpha'$, where α' is adjusted to ensure that the overall Type I error rate is controlled at the nominal level α.

To determine α', let $a = \Phi^{-1}(p_c)$. The conditional probability of rejecting H_0 given that the futility boundary is not crossed is:

$$P(T < c \mid T_1 < a) = \frac{1}{\Phi(a)} \int_{-\infty}^{a} \Phi(\sqrt{2}c - x)\, \phi(x)\, dx,$$

where ϕ is the standard normal density function. To ensure the overall Type I error rate is α, we solve for c in the equation:

$$p_c \cdot P(T < c \mid T_1 < a) = \alpha,$$

 doi:10.20944/preprints202507.1553.v1

and define the adjusted significance level as $\alpha' = \Phi(c)$, which satisfies $\alpha' > \alpha$. Once α' is determined numerically power can be calculated under H_1 via simulation. For a fixed n we can find combinations of n_1 and n_2 such that the overall $\alpha = 0.05$ and power is greater than or equal to the desired power. Practically speaking one can start at the n determined by the Simon two-stage design and reduce by increments of 1 until the power constraint is no longer satisfied. The search is over values of n_1 and p_c, with $n_2 = n - n_1$. This will be illustrated in the next section.

3.1. Convolution Approach and Simon Two-Stage Design Comparison

There is no straightforward way to directly compare the convolution-based approach with Simon's two-stage design. A key distinguishing feature is that the convolution-based method precisely controls the Type I error rate, whereas Simon's two-stage design may be conservative in some scenarios, as illustrated earlier in Figure 2. In settings where the actual Type I error of Simon's design falls well below the nominal level, the convolution-based method can achieve the same desired power with a smaller sample size, particularly when the difference between π_1 and π_0 is substantial.

To illustrate, we consider testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$ with $\pi_0 = 0.1$ and π_1 ranging from 0.3 to 0.6 in increments of 0.1. The results of various Simon two-stage designs are presented in Table 4, showing the corresponding total sample size, exact Type I error, power, expected sample size (EN_0), and probability of early stopping.

Similarly, Table 5 displays results from the convolution-based two-stage designs across the same values of π_1. For each design, we considered a range of p_c values from 0.2 to 0.7 in steps of 0.01, and we report the total sample size n, stage-wise sample sizes n_1 and n_2, p_c, power, expected sample size (ESN), adjusted α', for testing at the second stage, and the probability of early stopping, which is equal to $1 - p_c$. The designs in Table 5 represent a subset of possible scenarios to provide a concise summary.

The convolution-based approach demonstrates a clear advantage by achieving slightly smaller sample sizes across all settings. Moreover, it is not constrained by the range of early stopping probabilities observed in the Simon designs (approximately 0.549 to 0.810 in our example). Instead, the convolution-based method allows the user to specify any desired p_c (and thus early stopping probability), offering greater flexibility to tailor the design to the specific goals and constraints of a given clinical trial.

Table 4. Summary of Simon Two-Stage Designs with one-sided Type I error rate $\alpha = 0.05$, Power = 0.8, $\pi_0 = 0.1$. Designs are shown for various response probabilities π_1.

π_1	n	n_1	r_1	r_2	Type I Error	Power	EN_0	P(early stop)	Method
0.3	25	15	1	5	0.033	0.802	19.5	0.549	Minimax
0.3	26	12	1	5	0.036	0.805	16.8	0.659	
0.3	27	11	1	5	0.040	0.806	15.8	0.697	
0.3	29	10	1	5	0.047	0.805	15.0	0.736	Optimal
0.4	13	8	1	3	0.031	0.802	8.9	0.813	Minimax
0.4	15	4	0	3	0.043	0.818	7.8	0.656	Optimal
0.5	8	4	0	2	0.036	0.836	5.4	0.656	Minimax
0.5	9	3	0	2	0.041	0.828	4.6	0.729	Optimal
0.6	6	3	0	2	0.015	0.807	3.8	0.729	Minimax
0.6	8	2	0	2	0.025	0.819	3.1	0.810	Optimal

 doi:10.20944/preprints202507.1553.v1

Table 5. Summary of Convolution-Based Two-Stage Designs with one-sided Type I error rate $\alpha = 0.05$, Power = 0.8, $\pi_0 = 0.1$. Designs are shown for various response probabilities π_1.

π_1	n	n_1	n_2	p_c	Power	ESN	α'	P(early stop)
0.3	23	16	7	0.34	0.810	18.4	0.055	0.66
0.3	23	17	6	0.32	0.807	18.9	0.056	0.68
0.3	23	17	6	0.30	0.806	18.8	0.057	0.70
0.3	23	15	8	0.31	0.806	17.5	0.056	0.69
0.3	23	14	9	0.37	0.806	17.3	0.054	0.63
0.3	23	14	9	0.45	0.806	18.0	0.052	0.55
0.3	23	14	9	0.46	0.806	18.1	0.052	0.54
0.3	23	16	7	0.69	0.806	20.8	0.050	0.31
0.3	23	16	7	0.36	0.805	18.5	0.054	0.64
0.3	23	15	8	0.37	0.805	18.0	0.054	0.63
0.3	23	13	10	0.50	0.805	18.0	0.051	0.50
0.3	23	13	10	0.53	0.805	18.3	0.051	0.47
0.3	23	12	11	0.70	0.805	19.7	0.050	0.30
0.4	11	8	3	0.20	0.801	8.6	0.066	0.80
0.4	11	9	2	0.21	0.801	9.4	0.064	0.79
0.5	7	5	2	0.20	0.809	5.4	0.066	0.80
0.5	7	5	2	0.25	0.805	5.5	0.060	0.75
0.5	7	5	2	0.30	0.801	5.6	0.057	0.70
0.6	5	3	2	0.29	0.810	3.6	0.057	0.71
0.6	5	3	2	0.38	0.804	3.8	0.053	0.62
0.6	5	3	2	0.40	0.801	3.8	0.053	0.60
0.6	5	3	2	0.52	0.801	4.0	0.051	0.48

4. Real World Examples

4.1. One-Stage Designs

4.1.1. Example 1

Our first example [11] is from a study of patients with concomitant advanced non-small cell lung cancer (NSCLC) and interstitial lung disease (ILD). This prospective, multicenter, single-arm phase 2 trial investigated the efficacy and safety of albumin-bound paclitaxel (nab-paclitaxel) in combination with carboplatin in patients with both advanced NSCLC and ILD. The primary endpoint was the overall response rate (ORR), testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, with $\pi_0 = 0.2$, alternative $\pi_1 = 0.4$, $\alpha = 0.05$, and power = 0.80. Based on an exact binomial test, this required a sample size of $n = 35$. The actual study enrolled $n = 36$ subjects between April 2014 and September 2017, corresponding to an average accrual rate of approximately 1.3 subjects per month.

Using the same design parameters, the convolution-based test would require $n = 32$ subjects to achieve a power of 0.8117, or $n = 31$ for a power of 0.7967, potentially reducing the trial duration by 4 to 5 months with the corresponding cost savings. The p-value was < 0.001 under both the exact binomial and convolution-based approaches. If a prorated response of 17 out of 31 positive responses were observed, the p-value using the convolution-based method would still be < 0.001.

4.1.2. Example 2

Our next example study[12] enrolled $n = 15$ metastatic prostate cancer patients with AR-V7-expressing circulating tumor cells into a prospective phase II trial. The primary endpoint was PSA response, with hypothesis testing conducted as $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, where $\pi_0 = 0.05$, $\pi_1 = 0.264$, $\alpha = 0.10$, and target power = 0.80. The true Type I error rate for the exact binomial test was 0.0362. A positive outcome in the study was thus defined as ≥ 3 of 15 patients achieving a PSA response.

 doi:10.20944/preprints202507.1553.v1

Using the same design parameters, the convolution-based test would require $n = 11$ subjects to achieve a power of 0.824, or $n = 10$ for a power of 0.793. In the actual trial, 2 out of 15 subjects achieved a PSA response, yielding a p-value of 0.14 using the exact binomial test. The convolution-based test yielded a p-value of 0.106. Although not statistically significant at the $\alpha = 0.10$ level, this result demonstrates the relative efficiency of the convolution-based approach.

4.2. Two-Stage Designs

4.2.1. Example 1

Our next example[13] is based on a study evaluating vinorelbine in advanced non-small cell lung cancer (NSCLC) patients aged 70 years or older. The study employed a multicenter, two-stage phase II design following Simon's optimal method. The primary endpoint was objective response rate (ORR), with hypothesis testing structured as $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, where $\pi_0 = 0.10$, $\pi_1 = 0.25$, $\alpha = 0.05$, and target power = 0.80.

The final sample size under Simon's design was $n = 43$ with an interim futility analysis planned at $n_1 = 18$ subjects. The actual Type I error rate achieved was 0.048. Table 6 presents three example two-stage designs generated using the convolution-based estimator. Notably, if one is willing to delay the interim futility analysis beyond $n_1 = 18$, the total required sample size can be reduced from $n = 43$ to $n = 35$, offering a more efficient design compared to Simon's optimal design alternative. Even maintaining the interim analysis at $n_1 = 18$, the convolution-based approach can still reduce the total sample size to $n = 37$.

Table 6. Convolution-based design scenarios for Example 1: Two-stage design

n	n_1	n_2	π_c	Power	ESN	α'
35	27	8	0.23	0.803	28.8	0.062
36	24	12	0.24	0.802	26.9	0.061
37	18	19	0.56	0.801	28.6	0.051

In the observed trial data, 5 of the initial 18 subjects achieved a positive ORR, with a total of 10 ORR responses observed out of 43 by the end of the study. Simon's optimal design specified early stopping for futility if 3 or fewer ORR responses were observed in the first stage, and declared efficacy if 8 or more total responses were observed at the second stage.

For each convolution-based design in Table 6, we retrospectively aligned the observed data to the alternative designs to estimate what would have occurred under those configurations:

- For $n = 35$, $n_1 = 27$: Assuming 7 out of 27 responses, the futility p-value was 0.011, below the stopping threshold $\pi_c = 0.23$. The final-stage p-value, assuming 8 out of 35 total responses, was 0.0158, significant at the adjusted level $\alpha' = 0.062$.
- For $n = 36$, $n_1 = 24$: Assuming 6 out of 24 responses, the futility p-value was 0.002, below $\pi_c = 0.24$. The final-stage p-value based on 8 of 36 responses was 0.0162, also significant at $\alpha' = 0.061$.
- For $n = 37$, $n_1 = 18$: Assuming 5 out of 18 responses, the futility p-value was 0.015, which is less than $\pi_c = 0.56$. The final p-value with 8 of 37 responses was 0.020, significant at $\alpha' = 0.051$.

This example again highlights the potential cost savings and efficiency gains achievable with the convolution-based approach compared to Simon's two-stage design, while maintaining rigorous Type I error control and statistical power.

4.2.2. Example 2

Our next example[14] comes from a publication describing the LuDO-N trial, a phase II, open-label, multi-center, single-arm, two-stage clinical trial in children with high-risk neuroblastoma, utilizing an alternative administration schedule of Lutetium DOTATATE. The primary endpoint is the response rate, assessed by the Revised International Neuroblastoma Response Criteria one month after

 doi:10.20944/preprints202507.1553.v1

completion of therapy. Hypothesis testing is structured as $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$, where $\pi_0 = 0.2$, $\pi_1 = 0.4$, $\alpha = 0.1$, and the target power is 0.80.

The proposed design follows Simon's Two-Stage Minimax design. Recruitment is expected to be completed within 3–5 years. Based on the specified parameters, the Simon design requires a total sample size of $n = 24$, with a first-stage futility analysis at $n_1 = 14$, yielding a true Type I error rate of 0.0874.

In contrast, an alternative convolution-based design would require a total of $n = 19$ subjects, with $n_1 = 16$ in the first stage, using a futility threshold of $p_c = 0.21$ and a final adjusted significance level of $\alpha' = 0.158$. If the projected recruitment rate is 24 subjects over 5 years (approximately 4.8 subjects per year), the convolution-based design would reduce the total accrual period by roughly one year.

5. Conclusions

The convolution-based approach presented in this work offers a flexible and efficient alternative to traditional exact methods for designing and analyzing single-arm phase II oncology trials with binary endpoints. By convolving the binomial distribution with a simulated normal random variable, this method produces an unbiased estimator for the response rate π and achieves precise Type I error control. This leads to reduced sample size requirements in both one-stage and two-stage designs, while maintaining desirable operating characteristics.

A significant advantage of the convolution framework is its adaptability. It can be easily modified to incorporate an interim analysis for either futility or combined efficacy and futility, allowing for early decision-making and further optimization of trial resources. Moreover, the convolution-based method provides a smooth and continuous approximation to the binomial tail distribution, making it especially valuable in scenarios where measurement error or process variability exists, and a continuous p-value function is preferred.

Overall, this approach enhances the efficiency of early-phase oncology clinical trial design and provides a theoretically sound and practically implementable alternative to exact binomial and Simon's two-stage methods. It holds particular promise for high-cost therapeutic areas, such as oncology and cellular therapies, where efficient trial execution is critical.

Acknowledgments: This work was supported by the following three NCI grants to Dr. Hutson: nRG Oncology Statistical and Data Management Center grant (grant no. U10CA180822); Immuno-Oncology Translational network (IOTN) Moonshot grant (grant no. U24CA232979); Acquired Resistance to Therapy network (ARTNet) grant (grant no.U24CA274159). No potential competing interests were reported by the authors.

References

1. Prepared by Battelle Technology Partnership Practice. Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies. Prepared for Pharmaceutical Research and Manufacturers of America (PhRMA). 2015.
2. Leighl, N.B., Nirmalakumar, S., Ezeife, D. A. and Gyawali, B. (2021) An Arm and a Leg: The Rising Cost of Cancer Drugs and Impact on Access. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting* **41** 1-12.
3. Kapinos, K. A., Hu, E., Trivedi, J., Geethakumari, P. R. and Kansagra, A. (2023) Cost-Effectiveness Analysis of CAR T-Cell Therapies vs Antibody Drug Conjugates for Patients with Advanced Multiple Myeloma. *Cancer Control* **30** 1-8.
4. Hoover, A., Reimche, P., Watson, D., Tanner, L., Gilchrist, L., Finch, M., Messinger, Y. H. and Turcotte, L. M. (2024) Healthcare cost and utilization for chimeric antigen receptor (CAR) T-cell therapy in the treatment of pediatric acute lymphoblastic leukemia: A commercial insurance claims database analysis. *Cancer Reports*. Epub ahead of print https://doi.org/10.1080/19466315.2023.2261672
5. Simon, R. (1989) Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10** 1-10.
6. Hamaker, H. C. and Strik, R. van. (1955). The Efficiency of Double Sampling for Attributes. *Journal of the American Statistical Association* **50** 830–849
7. Duncan, A. J. (1986) *Quality Control and Industrial Statistics*. Irwin, Homewood, IL.

 doi:10.20944/preprints202507.1553.v1

8. Chernick, M. R. and Christine Y. Liu, C. Y. (2002) The Saw-Toothed Behavior of Power Versus Sample Size and Software Solutions: Single Binomial Proportion Using Exact Methods. *American Statistician* **56** 149–155.
9. Hutson, A. D. (2006) Modifying the Exact Test for a Binomial Proportion and Comparisons with Other Approaches. *Journal of Applied Statistics* **33** 679-690.
10. Whitlock, M. C.(2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* **18** 1368-73.
11. Asahina, H., Oizumi, S., Takamura, K., Harada, T., Harada, M., Yokouchi, H., Kanazawa, K., Fujita, Y., Kojima, T., Sugaya, F., Tanaka, H., Honda, R., Kikuchi, E., Ikari, T., Ogi, T., Shimizu, K., Suzuki, M., Konno, S., Dosaka-Akita, H., Isobe, H., Nishimura, M.; Hokkaido Lung Cancer Clinical Study Group. (2019) A prospective phase II study of carboplatin and nab-paclitaxel in patients with advanced non-small cell lung cancer and concomitant interstitial lung disease (HOT1302). *Lung Cancer* **138** 65-71.
12. Boudadi, K., Suzman, D. L., Anagnostou, V., Fu, W., Luber, B., Wang, H., Niknafs, N., White, J. R., Silberstein, J. L., Sullivan, R., Dowling, D., Harb, R., Nirschl, T. R., Veeneman, B. A., Tomlins, S. A., Wang, Y., Jendrisak, A., Graf, R. P., Dittamore, R., Carducci, M. A., Eisenberger, M. A., Haffner, M. C., Meeker, A. K., Eshleman, J. R., Luo, J., Velculescu, V. E., Drake, C. G. and Antonarakis, E. S. (2018) Ipilimumab plus nivolumab and DNA-repair defects in AR-V7-expressing metastatic prostate cancer. *Oncotarget* **9** 28561-28571
13. Gridelli, C., Perrone, F., Gallo, C., De Marinis, F., Ianniello, G., Cigolari, S., Cariello, S., Di Costanzo, F., D'Aprile, M., Rossi, A., Migliorino, R., Bartolucci, R., Bianco, A. R., Pergola, M. and Monfardini, S. (1997) Vinorelbine is well tolerated and active in the treatment of elderly patients with advanced non-small cell lung cancer. A two-stage phase II study. *European Journal of Cancer* bf 33 392-397.
14. Sundquist, F., Georgantzi, K., Jarvis, K. B., Brok, J., Koskenvuo, M., Rascon, J., van Noesel, M., Grybäck, .P, Nilsson, J., Braat, A., Sundin, M., Wessman, S., Herold, N., Hjorth, L., Kogner, P., Granberg, D., Gaze, M. and Stenman, J. (2022) A Phase II Trial of a Personalized, Dose-Intense Administration Schedule of 177Lutetium-DOTATATE in Children With Primary Refractory or Relapsed High-Risk Neuroblastoma-LuDO-N. Frontiers in Pediatrics 10 836230