Received: 8 March 2021 | Revised: 28 September 2021 | Accepted: 30 September 2021

DOI: 10.1002/pst.2174

WILEY

MAIN PAPER

Practical and robust test for comparing binomial proportions in the randomized phase II setting

Kristopher Attwood[1] | **Soyun Park**[1,2] | **Alan D. Hutson**[1]

[1]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

[2]Department of Biostatistics, University at Buffalo, Buffalo, New York, USA

Correspondence
Kristopher Attwood, Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Elm & Carlton Streets, Buffalo, NY 14263, USA.
Email: kristopher.attwood@roswellpark.org

Funding information
Roswell Park Comprehensive Cancer Center and National Cancer Institute, Grant/Award Number: P30CA016056; Immuno-Oncology Translation Network, Data Management and Resource-Sharing Center, Grant/Award Number: U24CA232979; NRG Oncology Statistical and Data Management Center, Grant/Award Number: U10CA180822

Abstract

The one-arm, non-randomized, one/two-stage phase II designs have been a mainstay in oncology trials for evaluating response rates or similar variants (i.e., tests about single proportions). With the goal of screening new therapies that have the potential to move into a randomized phase III trial or a subsequent randomized phase II trial, all while maintaining a logistically feasible sample size. However, since the implementation of the Food and Drug Administration's Fast Track Designation, there has been a trend toward randomized phase II clinical trials as a source of stronger evidence for those seeking fast-track approvals. While there are many single- and multi-stage randomized designs for evaluating proportions in this phase II setting, there still exist limitations in terms of sample size (which directly impacts cost and study duration) or operating characteristics (ex. maintained type I error). In this article, we propose a new test for comparing two binomial proportions, which is a modification across existing methods (the standard z-test and Jung's test). This approach is contrasted with existing methods via numeric evaluation and further contrasted using a real-world oncology trial. The proposed method demonstrates improvements in efficiency and robustness against deviations from design assumptions. When applied to the existing trial, significant savings with respect to cost and time are illustrated. Our proposed test for comparing binomial proportions provides an efficient and robust alternative in the randomized phase II oncology setting, especially when the control arm has a high rate.

KEYWORDS

clinical trial, cost effectiveness, exact testing, Fisher's exact test, phase II

1 | BACKGROUND

The one-arm non-randomized two-stage Simon design,[1] or similar variants,[2–5] have been a mainstay for oncology trials in the phase II setting for testing about a single proportion, for example, testing about the proportion of complete or partial responders. The goal of these designs is to screen new therapies that have the potential to move into a randomized phase III trial or a subsequent randomized phase II trial while maintaining a logistically feasible sample size.

The test of interest for the traditional single arm two-stage Simon design about a proportion takes the form:

$$H_0 : \pi = \pi_0 \text{ versus } H_1 : \pi > \pi_0,$$

where π_0 is the current response rate for the standard of care for a given cancer type, often determined by some criteria such as RECIST[6] or disease progression at some fixed time. The design itself has several desirable properties including an algorithm that allows one to search for the bounded α levels for each futility decision rule configuration near the desirable exact α rate. These designs treat π_0 as a fixed known value, though it should be noted that in many instances π_0 is wrongly treated as fixed when in fact it is derived from previous studies and should be considered a historical estimate.[7]

In recent years, investigators are turning to the two-arm randomized design for not only the statistical considerations (ex. bias, error control, and quality of evidence) and evolving treatment outcomes (ex. use of delayed tumor progression),[8] but also as an alternative business strategy. The Food and Drug Administration (FDA) has a fast-track designation,[9] which "provides for the designation of a drug as a fast-track product... if it is intended, whether alone or in combination with one or more other drugs, for the treatment of a serious or life-threatening disease or condition, and it demonstrates the potential to address unmet medical needs for such a disease or condition. This provision is intended to facilitate development and expedite review of drugs to treat serious and life-threatening conditions so that an approved product can reach the market expeditiously." Under this definition, many cancer disease site specific therapies are eligible for this designation. A randomized two-arm trial provides a stronger degree of evidence for approval in the fast track seeking as compared to a one-arm trial, providing strong motivation for carrying forth a randomized phase II trial.

However, a two-arm trial requires a substantially greater sample size as compared to a single-arm trial. For example, the maximum sample size needed for a Simon minimax two-stage design, stopping for futility, for testing $H_0 : \pi = 0.3$ versus $H_1 : \pi > 0.3$ to detect an alternative rate of $\pi = 0.5$ or larger at type I error rate $\alpha = 0.05$ and power $1-\beta = 0.8$ is $n = 39$. By comparison the single stage sample size for testing $H_0 : \pi_T = \pi_C$ versus $H_1 : \pi_T > \pi_C$ in a two-arm randomized setting, where π_T and π_C denote the new treatment and standard-of-care rates, respectively, requires a total sample size of $n = 83 + 83 = 166$ using Fisher's exact test at α=0.05 and $1 - \beta = 0.8$. This difference in sample size comes with an increased cost; where, in 2013, it was estimated that the average cost per subject in an oncology trial is roughly \$59,500.[10] Hence, under simplifying assumptions, the average approximate cost difference for our simple example is \$7,556,500.

In terms of design choices for testing $H_0 : \pi_T = \pi_C$ versus $H_1 : \pi_T > \pi_C$ in the randomized two-arm phase oncology II setting, the focus has been on generalizations of Simon type two-stage ideas for comparing two proportions using a variety of test statistics and methods for generating exact null distributions. If we let π_0 denote the historical standard-of-care rate that one would utilize in a typical one-arm design, Jung[11] developed an exact two-stage test about the null hypothesis $H_0 : \pi_T = \pi_C = \pi_0$ versus $H_1 : \pi_T > \pi_C, \pi_C = \pi_0$. The null distribution is then given by a straightforward product of two independent binomials under a fixed value for π_0, e.g., from our previous example above $\pi_0 = 0.3$. Decision rules for early stopping due to futility are obtained in a similar fashion to the approach of Simon[1] in the one-arm two-stage setting. The obvious limitation of this approach is again the reliance on the precise choice of π_0. This work was followed by an exact two-stage test based on Fisher's exact test[12] where in reference to the first test by Jung[11] it was noted that "...if the true response probabilities are different from the specified ones, the testing based on binomial distributions may not maintain the type I error close to the specified design value." Jung[11] proposed to use as a general rule $\pi_0 = 0.5$ for all designs. While this approach better maintains the type I error under misspecifications of the null, it also creates a very conservative test if the true $\pi_0 \neq 0.5$ and asymptotically converges to the incorrect distribution. A similar two-stage design is given by utilizing Barnard's test statistic.[13] Kepner also provides a general exact sequential procedure, which contains a two-stage design as a special case, stopping for futility, efficacy, or both.[14] Additional methods have been proposed in which the total sample size is minimized and the type I error rate controlled through critical region sculpting[15] or by utilization of a Bayesian framework.[16] In this note, we provide a method for modifying the single-stage version of Jung's approach[11] that is accurate, efficient, and robust to the choice of π_0.

In this work, we introduce our modified two-group test about proportions and compare the operating characteristics of our modified Jung test to commonly used methods (large sample z-test, Fisher's exact test, Barnard's unconditional test, and Jung's single-stage approach) via numeric evaluations and illustration involving an existing clinical trial. In the discussion, we further contrast our modified Jung test with the existing methods.

2 | TWO-GROUP TEST FOR COMPARING TWO PROPORTIONS

In keeping with the standard single-arm trials and Jung,[11] we have developed our test around a fixed historic rate (π_0) and the hypotheses:

$$H_0 : \pi_T = \pi_C = \pi_0 \text{ versus } H_1 : \pi_T > \pi_C = \pi_0. \quad (1)$$

Let X_T and X_C denote the number of responders from samples of size n_T and n_C, corresponding to the experimental treatment arm and the standard-of-care arm, respectively. The key feature of our approach is to incorporate an Agresti–Coull type continuity correction[17] that allows us to correct a z-type statistic to have precise and bounded type I error control. In addition, we will show that this approach is theoretically robust to misspecification of π_0. Toward this end, our modified Jung test statistic takes the form:

$$T(X_T, X_C; \delta) = \frac{\hat{\pi}_T - \hat{\pi}_C}{\sqrt{\frac{\hat{\pi}_T(1-\hat{\pi}_T)}{n_T+2} + \frac{\hat{\pi}_C(1-\hat{\pi}_C)}{n_C+2}}} + \frac{\delta}{\sqrt{n_T + n_C}}, \quad (2)$$

where $\hat{\pi}_T = \frac{X_T+1}{n_T+2}$, $\hat{\pi}_C = \frac{X_C+1}{n_C+2}$, and δ is a fixed constant depending upon α, π_0, n_T, and n_C. Adding one success and one failure to the estimators $\hat{\pi}_T$ and $\hat{\pi}_C$, similar to Agresti and Coull,[17] provides a mechanism such that we are not dividing by zero in Equation (2).

Determination of δ. First, let:

$$\hat{\alpha}_\delta = \sum_{i=0}^{n_T} \sum_{j=0}^{n_C} I_{(1-\Phi(T(X_T,X_C;\delta)) \le \alpha)} \times P(X_T = i|\pi_0) \times P(X_C = j|\pi_0), \quad (3)$$

where $I_{(\cdot)}$ denotes the indicator function, $P(X_T = i|\pi_0)$ is a $b(n_T, \pi_0)$ binomial probability calculated under H_0, $P(X_C = j|\pi_0)$ is a $b(n_C, \pi_0)$ binomial probability calculated under H_0, and $\Phi(\cdot)$ denotes the standard normal c.d.f. Now, determine the value of δ, which we will denote δ_0, such that $\min_\delta |\hat{\alpha}_\delta - \alpha|$ subject to the constraint $\hat{\alpha}_\delta \le \alpha$. The determination of δ_0 is a straightforward process from a numerical solutions standpoint. In essence, we are correcting the z test statistic in Equation (2) to have a precise and bounded α level in the finite sample setting using exact binomial probabilities calculated under H_0. For *unbalanced designs*, a minimum (n_{min}, where n_C and $n_T \ge n_{min}$) and maximum ($n = n_C + n_T$) sample sizes are specified, then all combinations of n_C, n_T, and δ_0 that satisfy the α constraint are identified. The optimum design is that combination which maximizes the power for the effect size of interest ($\Delta = \pi_1 - \pi_0$; where π_1 is a clinically relevant treatment rate).

Once δ_0 has been determined, the calculation of the one-sided p-value for testing the hypotheses specified in Equation (1) is given by $P = (1 - \Phi(T(X_T, X_C; \delta_0)))$; where we replace δ_0 for δ in Equation (2). A key feature of our approach that is distinct from Jung[11] is that the form of our test statistic provides an asymptotically type I error controlled and exact test for either a correctly specified π_0 or for $\pi_0 \in (l, u)$; where the value of δ_0 is constant for a given l and u under a pre-specified set of design conditions. For example, for $\alpha = 0.10$ and $n = 10 + 10 = 20$, we generated a plot of $\frac{\delta}{\sqrt{n_T+n_C}}$ versus π_0 in Figure 1.

As the sample sizes n_T and n_C increase jointly, the limit of the adjustment factor, $\frac{\delta}{\sqrt{n_T+n_C}}$, will tend toward 0 (as illustrated in Figure 2). The other component of the test statistic, $\frac{\hat{\pi}_T - \hat{\pi}_C}{\sqrt{\frac{\hat{\pi}_T(1-\hat{\pi}_T)}{n_T+2} + \frac{\hat{\pi}_C(1-\hat{\pi}_C)}{n_C+2}}}$, is asymptotically normally distributed. Therefore, in large samples, Equation (2) approaches normality and the type I error is asymptotically controlled. For large sample sizes, the exact approach and the asymptotic z-test approach have very similar type I error control.

An additional benefit to this approach is that unbalanced designs ($n_T \ne n_C$) are permitted without altering the test statistic. For a fixed $n = n_T + n_C$; the values of n_T, n_C, and δ would be selected such that the power is maximized, that is:

$$\max_{n_T, n_C, \delta} \sum_{i=0}^{n_T} \sum_{j=0}^{n_C} I_{(1-\Phi(T(X_T,X_C;\delta)) \le \alpha)} \times P(X_T = i|\pi_1) \times P(X_C = j|\pi_0),$$

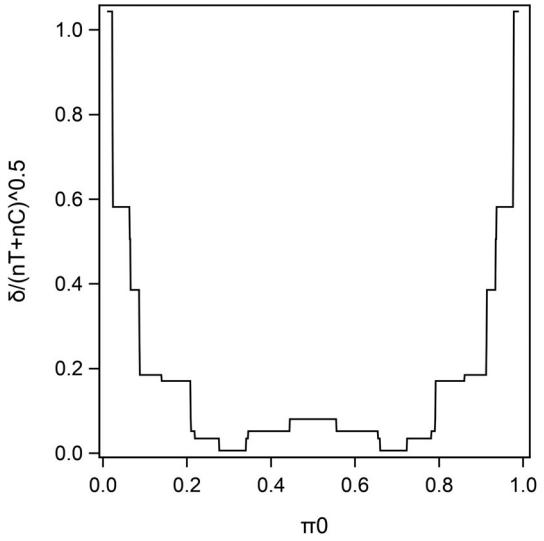

FIGURE 1 Plot of $\frac{\delta}{\sqrt{n_T+n_C}}$ versus π_0 for $\alpha = 0.05$ given a sample size of $n_T = n_C = 10$

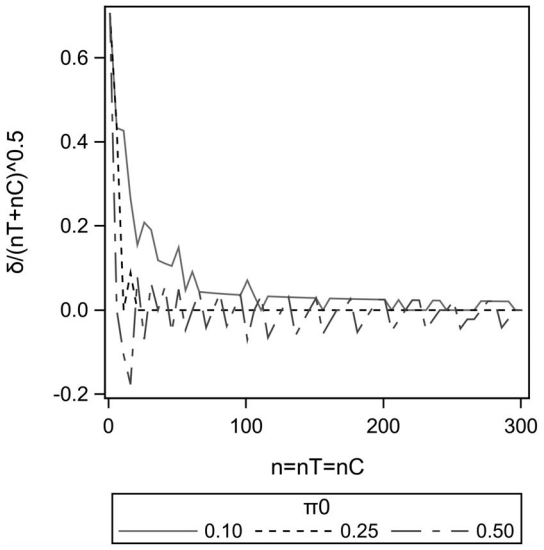

FIGURE 2 Plot of $\frac{\delta}{\sqrt{n_T+n_C)}}$ versus $n = n_T = n_C$ for $\alpha = 0.05$ and $\pi_0 = 0.1$, 0.25, and 0.5

where $\pi_1 > \pi_0$ is the treatment rate under the alternative hypothesis. In some scenarios, given a maximum total sample size, an unbalanced design can provide more power than the corresponding balanced design $(n_T = n_C)$. This is illustrated in Figure 3, where the power for testing $H_0 : \pi_T = \pi_C = 0.1$ versus $H_1 : \pi_T > \pi_C = 0.1$ across different effect sizes is plotted for unbalanced and balanced designs when $n_T + n_C = 40$.

3 | NUMERICAL EVALUATION OF COMPETING APPROACHES

The operating characteristics (type I error rate and power) of the modified Jung test (balanced and unbalanced designs) were evaluated numerically and compared to those of Fisher's exact test, Barnard's unconditional test (single stage version of Reference 13 and calculations made using the *Exact* package in R v4.0.2[18] with the *Boshloo* method for finding extreme tables), the asymptotic *z*-test, and the single-stage test of Jung.[10] Using binomial enumeration, the probability of rejection was calculated for the following hypotheses: $H_0 : \pi_T = \pi_C = \pi_0$ versus

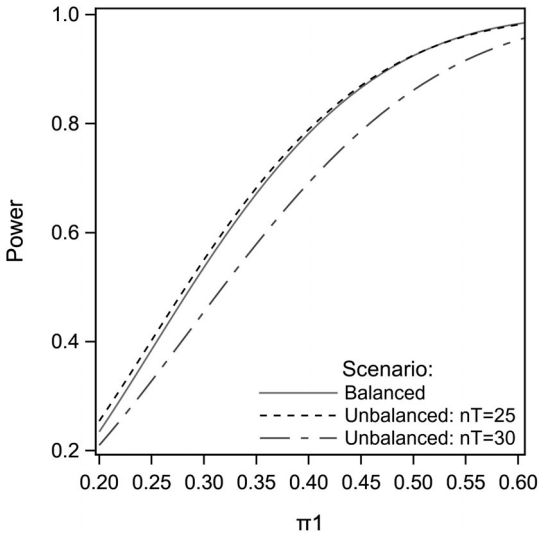

FIGURE 3 Plot of power versus π_T for balanced and unbalanced designs

$H_1 : \pi_T > \pi_C = \pi_0$. A range of realistic π_0's, effect sizes ($\Delta = \pi_T - \pi_C$), and sample sizes ($n = n_1 = n_2$) were considered, which are outlined in Table 1.

Additional numeric evaluations were used to evaluate the impact of an incorrectly specified π_0 on the type I error rates. Only the Jung type tests are evaluated, as the decision threshold for Jung's single-stage test and the test statistic (specifically δ) for the modified Jung test depend on π_0; whereas the null distributions of the test statistic corresponding to Fisher's exact and the z-tests are independent of π_0. Using binomial enumeration, the type I error rates were calculated for a variety of scenarios (outlined in Table 2) where the true and null π_0 are unequal.

The last numeric evaluations examine the efficiency of these tests in the design phase. For each test, the required sample size to achieve a minimum power (80% or 90%) is calculated for realistic significance levels (5% or 10%) and effect sizes (0.1 and 0.2). Additionally, using binomial enumeration, the actual power and type I error rates are calculated.

3.1 | Correctly specified π_0

Table 1 reports the numeric evaluation of power under a correctly specified π_0. In general, when π_0 is correctly specified, the type I error rates of the single-stage and modified Jung's tests are bounded by the target α; where Jung's single-stage test tended to have lower error rates. Within the sample size range we considered, as n increases, the type I error rate for the modified Jung test approaches the target α; whereas this is not the case for Jung's single-stage test. Fisher's exact test is also bounded, but the actual type I error rate tends to be much lower than specified (and relative to the other tests) and requires a large n to approach the specified α. Barnard's test is both bounded and approaches the target α as n increases. The z-test is not bounded by the target α and tends to have higher error rates relative to the Jung type tests, but does approach the target α for large n.

In terms of power, Jung's single-stage test tends to be the most powerful test when $\pi_0 < 0.5$ (noticeably when $\pi_C = 0.1$); while the modified Jung test and z-test tend to be the most powerful when $\pi_0 \geq 0.5$. This result may be due to the nature of the test statistics; where the test statistics for modified Jung test and z-test have symmetric distributions, but the distribution of Jung's single-stage test statistic is discrete, skewed, and bounded by n. Fisher's exact test had the lowest power across all examined scenarios, which is consistent with the significantly lower type I error rates. The modified Jung test was comparable to Barnard's test across all scenarios.

3.2 | Incorrectly specified π_0

Table 2 reports the numeric evaluation of power under an incorrectly specified π_0. When π_0 is incorrectly specified, the actual type I error rate for Jung's single-stage test can be significantly higher or lower than the target α.[11] These changes

TABLE 1 Numeric evaluation of power under a correctly specified π_0

$\alpha = 0.05$		$n = 40$					$n = 80$					$n = 120$				
π_C	π_T	F	Z	J	B	mJ	F	Z	J	B	mJ	F	Z	J	B	mJ
0.10	0.10	0.008	0.029	0.032	0.020	0.031	0.019	0.057	0.046	0.040	0.046	0.028	0.051	0.046	0.038	0.049
	0.15	0.041	0.096	0.113	0.068	0.109	0.093	0.186	0.196	0.150	0.175	0.144	0.206	0.243	0.179	0.207
	0.20	0.109	0.203	0.247	0.156	0.235	0.245	0.377	0.435	0.328	0.378	0.371	0.460	0.550	0.429	0.468
	0.25	0.213	0.337	0.413	0.279	0.385	0.450	0.583	0.673	0.536	0.597	0.633	0.712	0.805	0.688	0.725
	0.30	0.341	0.482	0.580	0.426	0.536	0.655	0.757	0.845	0.723	0.778	0.833	0.881	0.938	0.868	0.890
	0.40	0.619	0.746	0.836	0.707	0.671	0.915	0.949	0.980	0.941	0.959	0.984	0.991	0.997	0.989	0.992
0.25	0.25	0.025	0.054	0.049	0.042	0.047	0.031	0.052	0.046	0.047	0.048	0.032	0.052	0.036	0.044	0.048
	0.30	0.056	0.110	0.106	0.085	0.098	0.085	0.133	0.129	0.123	0.127	0.107	0.154	0.130	0.139	0.146
	0.35	0.106	0.194	0.193	0.149	0.171	0.184	0.268	0.270	0.247	0.258	0.255	0.329	0.309	0.314	0.320
	0.40	0.177	0.302	0.305	0.234	0.265	0.327	0.440	0.454	0.408	0.425	0.469	0.551	0.541	0.540	0.545
	0.50	0.375	0.554	0.572	0.459	0.490	0.666	0.767	0.799	0.742	0.753	0.859	0.898	0.897	0.888	0.897
0.50	0.50	0.021	0.059	0.040	0.041	0.047	0.028	0.047	0.046	0.046	0.048	0.041	0.060	0.041	0.042	0.050
	0.55	0.044	0.107	0.076	0.077	0.088	0.072	0.110	0.108	0.109	0.113	0.117	0.157	0.117	0.119	0.138
	0.60	0.083	0.180	0.131	0.133	0.154	0.156	0.221	0.215	0.216	0.228	0.260	0.323	0.260	0.268	0.303
	0.65	0.148	0.282	0.210	0.215	0.250	0.289	0.384	0.368	0.370	0.397	0.465	0.539	0.465	0.486	0.527
	0.75	0.375	0.554	0.438	0.459	0.526	0.666	0.767	0.727	0.742	0.779	0.859	0.898	0.858	0.888	0.897
0.75	0.75	0.025	0.054	0.049	0.042	0.047	0.031	0.052	0.046	0.047	0.048	0.032	0.052	0.036	0.044	0.048
	0.80	0.054	0.105	0.091	0.085	0.091	0.089	0.138	0.113	0.123	0.124	0.116	0.164	0.114	0.147	0.159
	0.85	0.111	0.193	0.158	0.159	0.164	0.220	0.311	0.240	0.279	0.279	0.315	0.394	0.282	0.370	0.390
	0.90	0.213	0.337	0.259	0.279	0.282	0.450	0.583	0.438	0.536	0.536	0.633	0.712	0.548	0.688	0.706
	0.95	0.376	0.546	0.401	0.469	0.469	0.750	0.863	0.685	0.830	0.830	0.916	0.948	0.825	0.937	0.939

Note: $n = n_T + n_C$. F = fisher's exact test. Z = asymptotic z-test. J = single-stage Jung's test. B = Barnard's unconditional test. mJ = modified Jung test with balanced sample size.

TABLE 2 Numeric evaluation of type I error under an incorrectly specified π_0

$\alpha = 0.05$		n = 40		n = 80		n = 120	
Assumed π_0	**True π_0**	**J**	**mJ**	**J**	**mJ**	**J**	**mJ**
0.1	0.050	0.0067	0.0067	0.0110	0.0226	0.0110	0.0352
	0.075	0.0179	0.0178	0.0276	0.0350	0.0280	0.0457
	0.100	0.0315	0.0311	0.0458	0.0460	0.0464	0.0492
	0.125	0.0456	0.0439	0.0630	0.0541	0.0637	0.0503
	0.150	0.0589	0.0541	0.0784	0.0584	0.0791	0.0521
	0.175	0.0710	0.0608	0.0918	0.0609	0.0926	0.0543
	0.200	0.0818	0.0642	0.1034	0.0631	0.1042	0.0554
0.25	0.15	0.0225	0.0296	0.0206	0.0447	0.0148	0.0460
	0.20	0.0367	0.0388	0.0342	0.0445	0.0260	0.0493
	0.25	0.0493	0.0473	0.0462	0.0480	0.0363	0.0485
	0.30	0.0595	0.0510	0.0560	0.0521	0.0449	0.0481
	0.35	0.0672	0.0494	0.0635	0.0508	0.0517	0.0520
	0.40	0.0727	0.0457	0.0687	0.0474	0.0564	0.0564
0.5	0.40	0.0373	0.0546	0.0431	0.0557	0.0381	0.0551
	0.45	0.0396	0.0496	0.0456	0.0503	0.0404	0.0531
	0.50	0.0403	0.0473	0.0465	0.0483	0.0412	0.0499
	0.55	0.0396	0.0496	0.0456	0.0503	0.0404	0.0531
	0.60	0.0373	0.0546	0.0431	0.0557	0.0381	0.0551
	0.65	0.0334	0.0584	0.0390	0.0591	0.0343	0.0519
0.75	0.65	0.0672	0.0494	0.0635	0.0508	0.0517	0.0520
	0.70	0.0595	0.0510	0.0560	0.0521	0.0449	0.0481
	0.75	0.0493	0.0473	0.0462	0.0480	0.0363	0.0485
	0.80	0.0367	0.0388	0.0342	0.0445	0.0260	0.0493
	0.85	0.0225	0.0296	0.0206	0.0447	0.0148	0.0460

Note: $n = n_T = n_C$. J = single-stage Jung's test. mJ = modified Jung test with balanced sample size.

appear to be dependent on both the specified and true π_0, and, within the range of values considered, do not appear to be mitigated by an increased n. The type I error rate for the modified Jung test does, however, approach the target α for large n.

3.3 | Design phase

Table 3 provides the sample size calculations for a specified type I error rate, power, and effect size. When $\pi_0 < 0.50$, the modified Jung test was less efficient relative to Jung' single-stage test, requiring a 0% to 18.9% larger n. It was significantly more efficient than Fishers exact test, requiring a 5.0% to 20.3% smaller n. At $\pi_0 = 0.5$, the modified Jung test was the most efficient method, requiring sample sizes that are 3.0% to 16.7%, 5.8% to 18.0%, and 0% to 7.4% smaller than those required for Jung's single-stage test, Fisher's exact test, and the z-test, respectively.

When $\pi_0 > 0.50$, the modified Jung test is still more efficient than Jung's single-stage and Fisher's exact tests (reductions in n of 9.7% to 32.1% and 6.2% to 17.4%, respectively). These results, on conjunction with the controlled type I error rates (for a properly or poorly defined π_0), indicate that the modified Jung test is preferable over the other approaches for scenarios where $\pi_0 \geq 0.50$.

TABLE 3 Sample size calculations for a specified type I error rate, power, and effect size

	Requested:			Required n per group					Actual power					Actual α				
π_0	Δ	α	power	F	Z	J	B	mJ	F	Z	J	B	mJ	F	Z	J	B	mJ
0.1	0.1	0.05	0.80	173	153	134	164	151	0.8003	0.8002	0.8016	0.8022	0.8017	0.0341	0.0524	0.0415	0.0373	0.0496
			0.90	232	212	182	221	212	0.9010	0.9008	0.9020	0.9010	0.9002	0.0363	0.0506	0.0482	0.0128	0.0499
		0.10	0.80	131	111	97	120	112	0.8023	0.8012	0.8035	0.8007	0.8019	0.0691	0.1018	0.0932	0.0285	0.0958
			0.90	183	163	141	173	163	0.9014	0.9000	0.9004	0.9013	0.9001	0.0733	0.1024	0.0979	0.0672	0.0976
	0.2	0.05	0.80	56	46	37	50	44	0.8025	0.8052	0.8087	0.8078	0.8032	0.0266	0.0521	0.0397	0.0467	0.0436
			0.90	74	64	53	70	62	0.9015	0.9014	0.9003	0.9094	0.9001	0.0308	0.0513	0.0369	0.0442	0.0497
		0.10	0.80	44	33	31	39	35	0.8086	0.8073	0.8129	0.8091	0.8045	0.0506	0.1020	0.0673	0.0917	0.0817
			0.90	58	48	40	54	47	0.9009	0.9007	0.9041	0.9049	0.9018	0.0526	0.1043	0.0940	0.0886	0.0971
0.25	0.1	0.05	0.80	277	258	251	264	259	0.8011	0.8005	0.8004	0.8018	0.8001	0.0411	0.0502	0.0444	0.0430	0.0494
			0.90	377	357	337	362	358	0.9009	0.9004	0.9006	0.9002	0.9003	0.0421	0.0502	0.0499	0.0287	0.0496
		0.10	0.80	207	189	178	190	190	0.8002	0.8007	0.8023	0.8001	0.8021	0.0807	0.1012	0.0992	0.0159	0.0994
			0.90	296	272	271	281	272	0.9009	0.9001	0.9001	0.9004	0.9000	0.0848	0.1006	0.0902	0.0337	0.0992
	0.2	0.05	0.80	78	69	65	71	70	0.8029	0.8035	0.8018	0.8023	0.8037	0.0347	0.0520	0.0423	0.0062	0.0498
			0.90	106	94	87	100	94	0.9024	0.9016	0.9001	0.9021	0.9004	0.0362	0.0513	0.0479	0.0323	0.0487
		0.10	0.80	59	49	47	53	47	0.8033	0.8009	0.8063	0.8040	0.8015	0.0701	0.1055	0.0946	0.0261	0.0986
			0.90	81	74	68	75	74	0.9009	0.9021	0.9037	0.9043	0.9013	0.0709	0.1037	0.0987	0.0365	0.0995
0.5	0.1	0.05	0.80	321	297	309	309	297	0.8011	0.8012	0.8011	0.8011	0.8000	0.0448	0.0547	0.0495	0.0395	0.0497
			0.90	445	419	432	432	419	0.9008	0.9003	0.9005	0.9007	0.9003	0.0436	0.0522	0.0477	0.0413	0.0500
		0.10	0.80	236	223	236	224	214	0.8014	0.8002	0.8014	0.8017	0.8013	0.0909	0.1005	0.0909	0.0848	0.0990
			0.90	341	328	328	328	316	0.9003	0.9006	0.9006	0.9006	0.9011	0.0901	0.0988	0.0988	0.0841	0.0989
	0.2	0.05	0.80	84	72	78	75	72	0.8052	0.8036	0.8031	0.8007	0.8032	0.0378	0.0565	0.0462	0.0375	0.0465
			0.90	111	102	108	102	96	0.9001	0.9033	0.9012	0.9015	0.9009	0.0349	0.0535	0.0444	0.0434	0.0497
		0.10	0.80	61	54	60	54	50	0.8023	0.8067	0.8046	0.8045	0.8038	0.0654	0.1054	0.0853	0.0934	0.0994
			0.90	89	76	82	82	76	0.9029	0.9031	0.9004	0.9029	0.9031	0.0771	0.1118	0.0921	0.0834	0.0955
0.75	0.1	0.05	0.80	215	193	226	200	196	0.8019	0.8002	0.8001	0.8012	0.8017	0.0399	0.0505	0.0460	0.0444	0.0496
			0.90	289	269	300	277	271	0.9001	0.9003	0.9009	0.9010	0.9003	0.0408	0.0505	0.0494	0.0140	0.0495
		0.10	0.80	162	142	166	150	143	0.8016	0.8022	0.8007	0.8021	0.8002	0.0803	0.1003	0.0915	0.0728	0.0993
			0.90	226	207	236	213	207	0.9001	0.9009	0.9012	0.9006	0.9009	0.0837	0.0997	0.0919	0.0284	0.0998
	0.2	0.05	0.80	45	34	53	40	37	0.8067	0.8056	0.8103	0.8297	0.8007	0.0298	0.0550	0.0459	0.0453	0.0492
			0.90	57	48	68	53	51	0.9007	0.9019	0.9007	0.9040	0.9041	0.0326	0.0553	0.0459	0.0195	0.0476

TABLE 3 (Continued)

	Requested:			Required n per group					Actual power					Actual α				
π_0	Δ	α	power	F	Z	J	B	mJ	F	Z	J	B	mJ	F	Z	J	B	mJ
		0.10	0.80	34	26	41	30	29	0.8056	0.8043	0.8081	0.8035	0.8109	0.0622	0.1022	0.0799	0.0980	0.0929
			0.90	46	37	56	41	38	0.9040	0.9007	0.9045	0.9041	0.9022	0.0634	0.1060	0.0777	0.0960	0.0989

Note: Δ = effect size. F = fisher's exact test. Z = asymptotic *z*-test. J = single-stage Jung's test. B = Barnard's unconditional test. mJ = modified Jung test with balanced sample size.

The Jung type tests are generally more efficient than Barnard's unconditional test across all scenarios. The modified Jung test tends to achieve a type I error rate closer to the specified level as compared to Barnard's test, which may explain the improved efficiency.

4 | ILLUSTRATIVE EXAMPLE

As a real-world illustration of the modified Jung test, we consider a randomized phase II trial of the addition of Bevacizumab to chemotherapy in the treatment of acute myeloid leukemia.[19] In this study, the complete response (CR) rate was compared between the chemotherapy alone and chemotherapy plus Bevacizumab arms using Fisher's exact test. The historic CR rate for the chemotherapy alone arm was 55%, while they expected a 15% increase by adding Bevacizumab. The study design ($n = 85$ per arm) achieved only 72.1% power at a one-sided significance level of 0.10. The study enrolled a total of $n = 171$ evaluable subjects over 30 months and concluded that there was no statistically significant difference in the CR rate.

A study design based on the modified Jung test would require only $n = 72$ subjects per arm and enrollment could be completed in approximately 25.3 months (assuming similar enrollment rates); a significant savings in both cost (approximately $1,547,000 in savings based on recent cost estimates[9]) and time. The study designs based on Jung's single-stage or Barnard's tests would require only $n = 80$ or $n = 73$ subjects per arm, respectively. An efficiency over Fisher's exact test, but less efficient than that of the modified Jung test.

The observed CR rate for the chemotherapy alone arm was 0.65, larger than the 0.55 specified in the design phase. Therefore, the actual type I error rate associated with Jung's design is not as specified in the design phase: design $\alpha = 0.088$ versus actual $\alpha = 0.076$. For a study design using the modified Jung test, the actual type I error rate ($\alpha = 0.095$) is nearly equivalent to that of the design phase ($\alpha = 0.099$) and closer to the target α of 0.10.

5 | DISCUSSION

In a time where the FDA fast track designation may lead to more randomized phase II studies, it is important to develop efficient, yet robust, statistical tools to control study costs and timelines. This article presents a novel and efficient test that evaluates the following hypotheses about proportions: $H_0 : \pi_T = \pi_C = \pi_0$ versus $H_1 : \pi_T > \pi_C, \pi_C = \pi_0$; which are in line with the traditional single-arm phase II trials.[1] The Jung type tests were generally more efficient and powerful than the standard Fisher's exact test, and comparable to Barnard's unconditional exact test. As compared to Barnard's test, this approach does rely on an additional assumption regarding π_0. However, we demonstrated that minor misspecification of π_0 does not have a dramatic impact on the operating characteristics of the proposed test. The single-stage Jung test[11] was the most efficient test when $\pi_0 < 0.5$; however, there is potential for inflated type I errors when π_0 is incorrectly specified. Our modified Jung test was the most efficient when $\pi_0 \geq 0.5$, which may provide this test with a niche in the setting of evaluating modifications to already effective treatments (e.g., adding immunotherapy to an existing treatment regimen). Additionally, this approach utilized a test statistic such that the type I error is effectively bounded, even under an incorrectly specified π_0. As with the Jung test,[11,12] this approach can be extended to multi-stage designs (ex. two-stage design) based on Simon type optimization criterion.[1] However, this becomes computationally intensive unless pre-specified futility criteria are utilized at the interim analysis, as in Reference 12.

In conclusion, the modified Jung test for comparing two binomial proportions would allow us to efficiently evaluate candidate treatments (especially when $\pi_0 \geq 0.5$) in a manner that is relatively robust to assumptions made in the design phase.

ACKNOWLEDGMENTS

This work was supported by Roswell Park Comprehensive Cancer Center and National Cancer Institute (NCI) grant P30CA016056, NRG Oncology Statistical and Data Management Center grant U10CA180822, and the Immuno-Oncology Translation Network: Data Management and Resource-Sharing Center at RPCI grant U24CA232979.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest for this work.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Kristopher Attwood https://orcid.org/0000-0002-7229-5472

REFERENCES

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10:1-10.
2. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982;38:143-151.
3. Kepner JL, Chang ML. Samples of exact k-stage group sequential designs for phase II and pilot studies. *Control Clin Trials*. 2004;25:326-333.
4. Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med*. 2004;23:561-569.
5. Lin Y, Shih WJ. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics*. 2004;60:482-490.
6. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228-247.
7. Chang MN, Shuster J, Kepner JL. Sample sizes based on exact unconditional tests for phase II clinical trials with historical controls. *J Biopharm Stat*. 2004;14:189-200.
8. Grayling MJ, Dimairo M, Mander AP, Jaki TF. A review of perspectives on the use of randomization in phase II oncology trials. *J Natl Cancer Inst*. 2019;111(12):1255-1262.
9. U.S. Dept. of HHS, FDA, CDER, and CBER. Guidance for Industry: Expedited Programs for Serious Conditions Drugs and Biologics Center for Drug Evaluation and Research, Silver Spring, MD. 2013.
10. 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.
11. Jung SH. Randomized phase II trials with a prospective control. *Stat Med*. 2008;27:568-583.
12. Jung SH, Sargent DJ. Randomized phase II clinical trials. *J Biopharm Stat*. 2014;24:802-816.
13. Shan G, Ma C, Hutson AD, Wilding GE. Randomized two-stage phase II clinical trial designs based on Barnard's exact test. *J Biopharm Stat*. 2013;23:1081-1090.
14. Kepner JL. On group sequential designs comparing two binomial proportions. *J Biopharm Stat*. 2010;20:145-159.
15. Litwin S, Basickes S, Ross EA. Two-sample binary phase 2 trials with low type I error and low sample size. *Stat Med*. 2017;36:1383-1394.
16. Cellamare M, Sambucini V. A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach. *Stat Med*. 2015;34:1059-1078.
17. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat*. 1998;52:119-126.
18. Calhoun P. Exact: Unconditional Exact Test. 2020. R package version 2.1.
19. Ossenkoppele GJ, Stussi G, Maertens J, et al. Addition of bevacizumab to chemotherapy in acute myeloid leukemia at older age: a randomized phase 2 trial of the Dutch-Belgian cooperative trial Group for Hemato-Oncology (HOVON) and the Swiss Group for Clinical Cancer Research (SAKK). *Blood*. 2012;120(24):4706.

How to cite this article: Attwood K, Park S, Hutson AD. Practical and robust test for comparing binomial proportions in the randomized phase II setting. *Pharmaceutical Statistics*. 2022;21(2):361-371. doi: 10.1002/pst.2174