A two-stage design to compare three treatments with binary responses

1. A two-stage comparison of two treatments

Two treatments T_1 and T_2 are to be compared, with one interim analysis. T_1 is the control treatment and T_2 is experimental. First, n_{i1} patients are to be randomised T_i, i = 1, 2, and then an interim analysis will be conducted. If necessary more patients will be randomised to T_i, bringing the total on that treatment to n_{i2}, i = 1, 2. Denote the success probability on T_i by p_i. Let the log odds-ratio be

$$\theta = log\left\{\frac{p_1(1-p_2)}{p_2(1-p_1)}\right\}.$$

At the h^th^ interim analysis, let the number of successes on T_i be S_{ih}, h = 1, 2; i = 1, 2. Define the test statistics

$$Z_h = \frac{n_{2h}S_{1h} - n_{1h}S_{2h}}{n_{1h} + n_{2h}}$$

and

$$V_h = \frac{n_{1h}n_{2h}(S_{1h} + S_{2h})(n_{1h} + n_{2h} - S_{1h} - S_{2h})}{(n_{1h} + n_{2h})^3}$$

h = 1, 2. If $Z_1/\surd V_1 \geq -0.6128$, then the trial will stop and T_2 will be discarded. The critical value corresponds to a one-sided nominal p-value of 0.27.

If neither treatment is declared superior, then a further n_{i2} patients will be randomised to T_i. If $Z_2/\surd V_2 \leq -c$, then T_2 will be declared superior. Otherwise, it will be concluded that there is no difference between the treatments.

Suppose that n patients receive the experimental treatment and 2n the control prior to the interim analysis, and that the same number of additional patients is treated if the trial continues beyond the interim analysis. Using the SAS program `OHSS two-stage - find c` shows that, for selection of T_2 to correspond to one-sided significance at the 2.5% level, c should be set at 1.92134. The program `OHSS two-stage - find n` shows that the power of the procedure when $p_1 = 0.7$, $p_2 = 0.9$ and n is fixed at 27 will be 0.917. Note that success probabilities (of avoidance of hospitalisation) are used here.

Simulating using `OHSS two-stage - sim.sas` evaluates this design. With $p_1 = 0.7$ and $p_2 = 0.9$, million-fold simulations give the probability of finding T_2 to be superior to be 0.850, with a probability of eliminating T_2 at the interim of 0.056 and an expected sample size of 157. With $p_1 = p_2 = 0.7$, million-fold simulations give the probability of finding T_2 to be superior to be 0.0242, with a probability of eliminating T_2 at the interim of 0.723 and an expected sample size of 103. The relatively small sample size, combined with scenarios in which the success probability can be close to 1, leads to situations where the standard normal approximation to the distribution of $Z/\surd V$ is only approximate. Thus, the power is found to be 0.85 rather than 0.917, and the null early stopping probability is 0.723 rather than 0.730.

With $p_1 = 0.7$ and $p_2 = 0.76$, million-fold simulations give the probability of finding T_2 to be superior to be 0.117, with a probability of eliminating T_2 at the interim of 0.512 and an expected sample size of 121. Such a difference would "lead to marked healthcare cost savings", and yet – were it to be true – it would be unlikely to be demonstrated in the trial.

2. A two-stage comparison of three treatments

The three treatments are denoted by T_1, T_2 and T_3, with respective success probabilities p_1, p_2 and p_3. T_1 is the control, and T_2 and T_3 are experimental treatments. The log odds-ratio between treatments T_i and T_j, is θ_{ij}, and the corresponding test statistics at the h^{th} analysis are Z_{ijh} and V_{ijh}, h = 1, 2. In the first stage of the study, n_{i1} patients are randomised to T_i, i = 1, 2, 3. If $Z_{1j1}/\sqrt{V_{1j1}} \geq -0.6128$, then T_j is eliminated, j = 2, 3. If two treatments are eliminated, then the trial stops. If only one treatment is eliminated, or if none are, then a second stage is conducted. This involves the randomisation of more patients to the remaining treatments, so that the total on T_i becomes n_{i2}. If $Z_{1j2}/\sqrt{V_{1j2}} \leq -c$, then T_i is declared to be superior to control, j = 2, 3.

We require the following two properties. *Type I error requirement*: For treatment T_i, if $p_j = p_1$, then the probability that the trial finds T_i to be superior to control is to be ≤ 0.025. *Power requirement*: If T_j is superior to T_1 to the extent that $\theta_{1j} = 1.350$ (corresponding to $p_1 = 0.7$ and $p_j = 0.9$), then the probability that T_j would be found superior to control is to be ≥ 0.90. The two-stage design will involve randomising 2n patients to control and n patients to each experimental treatment in the first round, and then 2n to control and n to any remaining experimental treatment if there is to be a second round, with n = 27 and c = 1.92134, as deduced for the two-treatment case in Section 1.

The program `two-stage 3 - sim` performs simulations of this design, and some results are shown in Table 1. The type I error requirement specifies that whenever $p_j = p_1$, choose_j should be ≤ 0.025. This is achieved, with the probability of such an error being estimated to be 0.024 in each case. The power requirement specifies that choose_j should be ≥ 0.90 whenever $p_1 = 0.7$ and $p_j = 0.90$. The value 0.85 is achieved, this value being the same as in the two-treatment case, with the inaccuracy due to the small sample size rather than the interim analysis. When $p_1 = 0.7$ and $p_j = 0.70$, the power is found to be 0.118, for the two-treatment case the corresponding value was 0.117. The quantity choose is the probability that one or both of the experimental treatments are found significantly better than control. When neither are, the risk of this error is 0.046, which might be acceptable. When both achieve the target improvement, then the probability if recommending at least one of then is 0.953, which is a desirable property. The last row depicts a situation in which T_3 achieves the target improvement but T_2 just misses it. The probability of at least one of these treatments being chosen is then 0.90. In this case, if the second best were the only one chosen, this outcome would still be quite desirable.

The program `estimation 3 - sim` performs simulations to determine the biases of naïve estimators applied at the end of the trial and the coverage probabilities of corresponding 95% confidence intervals. Some results are shown in Table 2. In general, the control success probability is overestimated and that of the experimental treatments underestimated. The biases are small, sometimes negligible. The log odds-ratio between 0.7 and 0.9 is –1.350 and between 0.7 and 0.76 it is –0.305. Zero differences between experimental treatments and control are overestimated while positive differences are underestimated. Estimated differences between the active treatments show little bias.

Table 1: Properties of the two-stage design deduced from million-fold simulation

E(N): average total number of patients recruited

stop: proportion of trials that stopped at the interim analysis

choose$_j$: proportion of trials in which T_j was found to be superior to control

Case	p_1	p_2	p_3	E(N)	stop	choose$_2$	choose$_3$	choose
1	0.70	0.70	0.70	146	0.566	0.024	0.024	0.046
2	0.70	0.70	0.90	192	0.051	0.024	0.850	0.851
3	0.70	0.90	0.90	212	0.011	0.850	0.850	0.953
4	0.70	0.70	0.76	160	0.419	0.024	0.118	0.134
5	0.70	0.76	0.76	171	0.322	0.118	0.118	0.206
6	0.70	0.85	0.90	208	0.024	0.556	0.850	0.900

Table 2: Properties of naïve estimators deduced from million-fold simulation

	True value of			Estimate of						Confidence interval coverage					
Case	p_1	p_2	p_3	p_1	p_2	p_3	θ_{12}	θ_{13}	θ_{23}	p_1	p_2	p_3	θ_{12}	θ_{13}	θ_{23}
1	0.70	0.70	0.70	0.708	0.688	0.688	0.088	0.088	0.000	0.942	0.950	0.950	0.951	0.951	0.944
2	0.70	0.70	0.90	0.702	0.688	0.897	0.088	–1.094	–1.313	0.947	0.950	0.910	0.951	0.941	0.974
3	0.70	0.90	0.90	0.701	0.897	0.897	–1.094	–1.094	0.000	0.951	0.910	0.910	0.942	0.942	0.941
4	0.70	0.70	0.76	0.708	0.688	0.747	0.088	–0.195	–0.325	0.940	0.950	0.950	0.951	0.956	0.942
5	0.70	0.76	0.76	0.708	0.747	0.747	–0.195	–0.195	0.000	0.940	0.950	0.950	0.956	0.956	0.939
6	0.70	0.85	0.90	0.701	0.843	0.897	–0.737	–1.094	–0.489	0.949	0.907	0.910	0.964	0.942	0.951

Some of the naïve confidence intervals for absolute success probabilities are too small, the lowest estimate of coverage being 0.907. Those for treatment differences are often conservative, the lowest value being 0.939. The naïve estimates and confidence intervals are computed using what Whitehead, Desai and Jaki (2020) refer to as Option 2. That is, only data collected when randomisation was open to both T_i and T_j are used to compute the estimate of and confidence interval for θ_{ij}. This contrasts with Option 1, in which all available data are used to compute all estimates.

There is an accuracy issue to be addressed, although it is quite minor. The most important discrepancies concern the biases in estimates of treatment difference.

3. Analysis of a three-treatment trial using Rao-Blackwellisation

Following Whitehead, Desai and Jaki (2020), the Rao-Blackwellisation approach will be used to analyse the completed data from the two-stage design. This avoids biases in estimates and inaccuracies in confidence interval coverage due to the potential for stopping the study at the interim analysis. Generally speaking, the estimate is based on the estimate $\hat{\theta}_1 = Z_1 / V_1$ deduced from the data available at the first interim analysis, which is unbiased for θ as it does not depend on the stopping rule in any way. Consequently, the estimate the estimate $\tilde{\theta} = E\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)$ is used, where **S*** and **n*** are the vectors of numbers of successes and numbers of patients, by treatment, in the final dataset. The final dataset might be that from the interim analysis or from the second analysis, depending whether early stopping occurs. Now $E\left(\tilde{\theta}\right) = \theta$, and

$$\operatorname{var}\left(\tilde{\theta}\right) = \operatorname{var}\left\{E\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)\right\} = \operatorname{var}\left(\hat{\theta}_1\right) - E\left\{\operatorname{var}\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)\right\} = \left(1/V_1\right) - E\left\{\operatorname{var}\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)\right\}.$$

In order to compute confidence intervals, it will be assumed that the pivot $\left\{\tilde{\theta} - E\left(\tilde{\theta}\right)\right\} \big/ \sqrt{\operatorname{var}\left(\tilde{\theta}\right)}$ follows a standard normal distribution and that $E\left\{\operatorname{var}\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)\right\}$ can be reliably estimated by $\operatorname{var}\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)$. Thus the standard error of $\tilde{\theta}$ is given by

$$\operatorname{se}\left(\tilde{\theta}\right) = \sqrt{\left\{\left(1/V_1\right) - \operatorname{var}\left(\hat{\theta}_1 \middle| \mathbf{S}^*, \mathbf{n}^*\right)\right\}} \ ,$$

and an approximate 95% confidence interval for θ is $\left(\tilde{\theta} \pm 1.96\operatorname{se}\left(\tilde{\theta}\right)\right)$.

In the case of a three-treatment trial yielding binary data, direct computation $\tilde{\theta}$ and its standard error are feasible. A similar approach can be taken to the estimation of the absolute success probabilities p_1, p_2 and p_3. It will be supposed that the sample sizes n_{ih} of patients receiving treatment T_i at the hth analysis are fixed in advance.

Table 3: Analyses of trial data

Treatment	T_1		T_2		T_3	
Data	n	S	n	S	n	S
Interim analysis	54	38	27	24	27	18
Final analysis	108	75	54	49	27	18

Method	$p_{1,lo}$	$p_{1,hat}$	$p_{1,hi}$	$p_{2,lo}$	$p_{2,hat}$	$p_{2,hi}$	$p_{3,lo}$	$p_{3,hat}$	$p_{3,hi}$
Data from interim analysis	0.582	0.704	0.826	0.770	0.889	1.007	0.489	0.667	0.844
Naïve analysis	0.608	0.694	0.781	0.830	0.907	0.985	0.489	0.667	0.844
RB analysis	0.606	0.696	0.786	0.818	0.908	0.998	0.489	0.667	0.844

Method	$\theta_{12,lo}$	$\theta_{12,hat}$	$\theta_{12,hi}$	$\theta_{13,lo}$	$\theta_{13,hat}$	$\theta_{13,hi}$	$\theta_{23,lo}$	$\theta_{23,hat}$	$\theta_{23,hi}$
Data from interim analysis	–2.122	–1.031	0.059	–0.827	0.174	1.174	0.003	1.286	2.569
Naïve analysis (Option 1)	–1.957	–1.186	–0.415	–0.781	0.130	1.041	0.462	1.684	2.906
RB analysis (Option 1)	–2.106	–1.190	–0.275	–0.768	0.147	1.061	0.373	1.466	2.560
Naïve analysis (Option 2)	–1.957	–1.186	–0.415	–0.827	0.174	1.174	0.003	1.286	2.569
RB analysis (Option 2)	–2.106	–1.190	–0.275	–0.827	0.174	1.174	0.003	1.286	2.569

Given that S_{i2} successes were observed on T_2 at the second analysis, then the value of S_{i1} is a hypergeometric random variable, being the number of red balls drawn in a sample of n_{i1} from an urn containing n_{i2} balls of which S_{i2} are red. Notice that, if T_i is eliminated at the interim analysis then $n_{i2} = n_{i1}$ and $S_{i2} = S_{i1}$ but the distributional statement remains trivially true: all of the balls will be drawn from the urn and S_{i1} of them will be certain to be red. This helps in the program, as no separate routine is needed when treatments are eliminated.

Table 3 presents an illustrative data set, and several alternative analyses, computed using `OHSS rbanalysis.sas`. At the interim analysis, T_3 is eliminated. Estimates and confidence intervals are given for the three success probabilities, first based on the data available at the interim analysis, second based on that from the final analysis, and third based on Rao-Blackwellisation. The RB adjustment makes T_1 and T_2 look very slightly better in the third decimal place, while not affecting T_3 at all. The lack of effect on T_3 is guaranteed, as no additional data are collected on that treatment. The 95% confidence intervals for p_1 and p_2 are widened by the RB process.

For treatment comparisons, two options are available. Either, treatments are compared using all data on each of them (Option 1), or they are compared using only the data collected during the time when both were available for randomisation (Option 2). Option 2 guards against time trends in success rates and means that the second set of data is not used when comparing T_3 with T_1 or T_2, but it is used when comparing T_1 with T_2. The adjusted comparison very slightly increases the estimate of the disadvantage of T_1 over T_2, and provides a wider confidence interval. When Option 1 is used, the estimated advantage of T_1 over T_3 is slightly increased, while that of the advantage of T_2 over T_3 is decreased. Confidence intervals are again widened, except for that of the advantage of T_2 over T_3 when using Option 1.

4. Evaluation through simulation

The two-stage design introduced in Section 2 has been evaluated using 10,000-fold simulations. The results are shown in Table 4. Five methods of analysis are investigated: use of the interim data only, use of naïve estimates that ignore the stopping and elimination rules and the Rao-Blackwellisation method of Section 3. The latter two methods are applied first using Option 1 and then using Option 2. Choice of option has no effect on the estimates of absolute probabilities. Computations use `OHSS rbsim.sas`.

In all cases, estimates of absolute success probabilities based on the interim data are virtually unbiased. Naive analyses of the complete data tend to overestimate the success rates on control (T_1) and underestimate success rates on active treatments, although the biases are small. Rao-Blackwellisation leads to estimates of absolute success rates based on the complete data that are as good as those from the interim data. Coverage probabilities of confidence intervals are too small in most cases, being closest to 0.95 for the naive analyses. The RB analyses lead to the worst coverage of all when the true success rate is 0.90. This could be due to poor approximations for a parameter close to the edge of its range, although there could still be programming errors to be found. For 10,000 replicates, the 95% probability interval for estimation of 0.95 is (0.946, 0.954).

Treatment differences of zero are well estimated from the interim data. Naive estimates introduce some bias, but RB analyses are able to remove it successfully.

However, non-zero treatment differences are poorly estimated bay all methods. Most commonly, this is underestimation of magnitude, although in some comparisons of T_2 and T_3 there is overestimation. In Cases 4 and 5 where the true treatment difference is modest, Rao-Blackwellisation provides a worthwhile reduction of bias of non-zero differences. For larger differences, although the RB analyses are less biased, the improvement is dwarfed by the bias due to small sample sizes being inadequate for distributional approximations. The coverage probabilities of most confidence intervals are adequate, and RB methods appear generally to increase them. However, there are cases where RB appears to make things worse, and the performance is especially poor when one or both treatments being compared has a success rate of 0.9.

5. Conclusion

The main difficulty with this trial design is that the planned sample sizes are very small. Power is intended to be 0.90 when treatments with success probabilities of 0.70 and 0.90 are compared. Such a large treatment effect is unusual, and it appears that the investigators are not really expecting to find it. The power to detect a treatment that improves the success rate from 0.70 to 0.76 is only 0.118, although that is the magnitude of improvement that is felt to be more realistic. Hence, the likelihood is that this trial will end with a trend for one or both experimental treatments to be better than control, but no significant evidence of advantage and no scientific basis for persuading practitioners to adopt the apparently better treatment. It will either prove to be a waste of time and resources or else the basis for unscientific adoption of a new treatment that might in the long-term be found to offer no improvement.

The main concern with post-trial estimation of treatment differences is that the small sample sizes lead to poor results for estimation methods based on assuming that test statistics are normally distributed. Nevertheless, this problem is less severe when the competing treatments have success rates between 0.70 and 0.76, which is likely to be the true situation. In that case, Rao-Blackwellisation does reduce bias and lead to reasonably accurate coverage rates for confidence intervals, and so appears to be worth applying.

Table 4: Evaluation of the two-stage design based on 1,000-fold simulations

	Mean estimate			Coverage probability			Mean estimate			Coverage probability		
Method	p_1	p_2	p_3	p_1	p_2	p_3	θ_{12}	θ_{13}	θ_{23}	θ_{12}	θ_{13}	θ_{23}
Case 1: True values	0.70	0.70	0.70				0.000	0.000	0.000			
Data from interim analysis	0.700	0.699	0.699	0.942	0.930	0.928	0.004	0.003	0.000	0.951	0.950	0.947
Naïve analysis (Option 1)	0.708	0.687	0.688	0.941	0.951	0.950	0.102	0.100	–0.001	0.947	0.946	0.958
RB analysis (Option 1)	0.700	0.699	0.699	0.936	0.923	0.924	0.003	0.003	–0.001	0.955	0.954	0.940
Naïve analysis (Option 2)							0.092	0.091	–0.003	0.950	0.949	0.945
RB analysis (Option 2)							0.003	0.002	–0.002	0.955	0.955	0.947
Case 2: True values	0.70	0.70	0.90				0.000	–1.350	–1.350			
Data from interim analysis	0.700	0.699	0.900	0.942	0.930	0.939	0.004	–1.126	–1.270	0.951	0.966	0.974
Naïve analysis (Option 1)	0.702	0.687	0.897	0.946	0.951	0.910	0.072	–1.094	–1.457	0.945	0.942	0.944
RB analysis (Option 1)	0.700	0.699	0.900	0.943	0.922	0.758	0.002	–1.126	–1.270	0.954	0.967	0.972
Naïve analysis (Option 2)							0.003	–1.126	–1.272	0.950	0.942	0.975
RB analysis (Option 2)							0.092	–1.094	–1.317	0.956	0.967	0.971
Case 3: True values	0.70	0.90	0.90				–1.350	–1.350	0.000			
Data from interim analysis	0.700	0.900	0.900	0.942	0.942	0.939	–1.124	–1.126	–0.004	0.967	0.966	0.946
Naïve analysis (Option 1)	0.701	0.897	0.897	0.950	0.912	0.910	–1.100	–1.101	–0.002	0.943	0.944	0.939
RB analysis (Option 1)	0.700	0.900	0.900	0.940	0.761	0.759	–1.125	–1.126	–0.002	0.963	0.962	0.876
Naïve analysis (Option 2)							–1.092	–1.094	–0.004	0.940	0.942	0.940
RB analysis (Option 2)							–1.125	–1.126	–0.003	0.968	0.967	0.880

Table 4 (continued)

	Mean estimate			Coverage probability			Mean estimate			Coverage probability		
Method	p_1	p_2	p_3	p_1	p_2	p_3	θ_{12}	θ_{13}	θ_{23}	θ_{12}	θ_{13}	θ_{23}
Case 4: True values	0.70	0.70	0.76				0.000	–0.305	–0.305			
Data from interim analysis	0.700	0.699	0.760	0.942	0.930	0.901	0.004	–0.299	–0.311	0.951	0.955	0.947
Naïve analysis (Option 1)	0.708	0.687	0.746	0.941	0.951	0.949	0.101	–0.191	–0.308	0.946	0.952	0.948
RB analysis (Option 1)	0.700	0.699	0.760	0.938	0.924	0.915	0.002	–0.300	–0.311	0.955	0.962	0.943
Naïve analysis (Option 2)							0.092	–0.194	–0.328	0.950	0.955	0.943
RB analysis (Option 2)							0.003	–0.300	–0.312	0.955	0.962	0.948
Case 5: True values	0.70	0.76	0.76				–0.305	–0.305	0.000			
Data from interim analysis	0.700	0.759	0.760	0.942	0.904	0.901	–0.298	–0.299	–0.001	0.956	0.955	0.945
Naïve analysis (Option 1)	0.708	0.747	0.746	0.940	0.951	0.949	–0.193	–0.193	0.001	0.951	0.951	0.950
RB analysis (Option 1)	0.700	0.760	0.760	0.940	0.917	0.915	–0.300	–0.299	0.001	0.963	0.963	0.940
Naïve analysis (Option 2)							–0.194	–0.194	–0.001	0.955	0.955	0.942
RB analysis (Option 2)							–0.300	–0.299	–0.001	0.964	0.963	0.948
Case 6: True values	0.70	0.85	0.90				–0.887	–1.350	–0.463			
Data from interim analysis	0.700	0.850	0.900	0.942	0.917	0.939	–0.806	–1.126	–0.467	0.969	0.966	0.960
Naïve analysis (Option 1)	0.701	0.843	0.897	0.948	0.908	0.910	–0.756	–1.098	–0.505	0.958	0.943	0.934
RB analysis (Option 1)	0.700	0.850	0.900	0.940	0.888	0.759	–0.807	–1.126	–0.464	0.971	0.964	0.926
Naïve analysis (Option 2)							–0.737	–1.094	–0.487	0.961	0.942	0.949
RB analysis (Option 2)							–0.807	–1.126	–0.463	0.972	0.967	0.932