Appendix: New Simulations

2020-10-09

Stopping with moderately permissive outcome (p1)

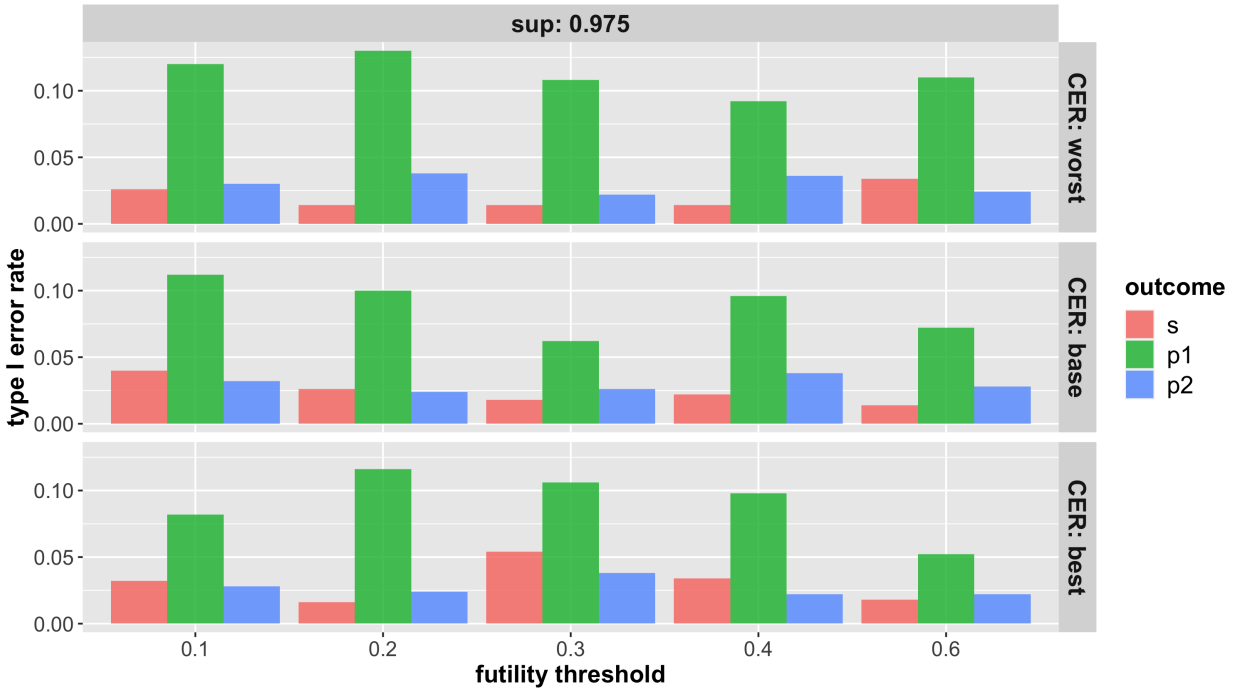
Maximum allowed sample size of 12,000 with interim analyses conducted every 1,000 patients

Both superiority and futility decisions were made with respect to the moderately permissive outcome (p1). Unless stated otherwise, the futility bound used was 0.4 and a probability threshold of 0.99 was required to stop for futility. The number of expected interim analyses can be thought of as the expected number of patients at trial termination divided by the batch size, which here is 1,000 patients.

Type I error (false positive risk)

Figure 1 displays the overall type I error rate for the simulations where early stopping is based on interim results from the moderately permissive outcome (p1). The type I error is under good control (i.e., <5%) for outcomes s and p2 since repeated superiority testing is only occurring for outcome p1. For outcome p1, the type I error is approximately 10% the superiority threshold 0.975.

Figure 1: Overall type I error rate based on moderately permissive (p1) outcome based stopping rules.Observed type one error rate for each outcome is presented by colored bars (see legend). Results by the three categories of control event rates are by rows. Results by the futility thresholds are presented on the x-axis.



Power

Figure 2 displays the power for each of the outcomes under each of the simulated scenarios and employed stopping rules. Generally, the power to detect an effect is high for outcome p2 in all scenarios (except RRR = 0.1). The power to detect an effect for outcome p1 is high in all scenarios where RRR=40% or 60%, but moderate to low when RRR is less than 20%. The power to detect an effect on outcome s is low in all cases, but achieves its highest estimates when RRR=60% and CER best case scenario.

Figure 2: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels). Power estimates by control event rates are presented by the rows. Power estimates by the futility thresholds are presented by the bar color (see legend).

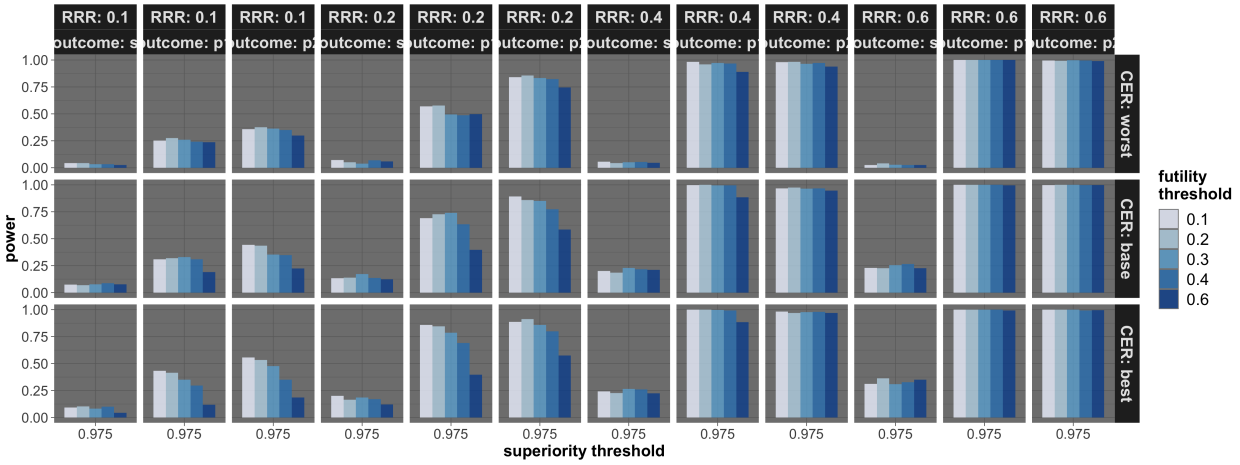
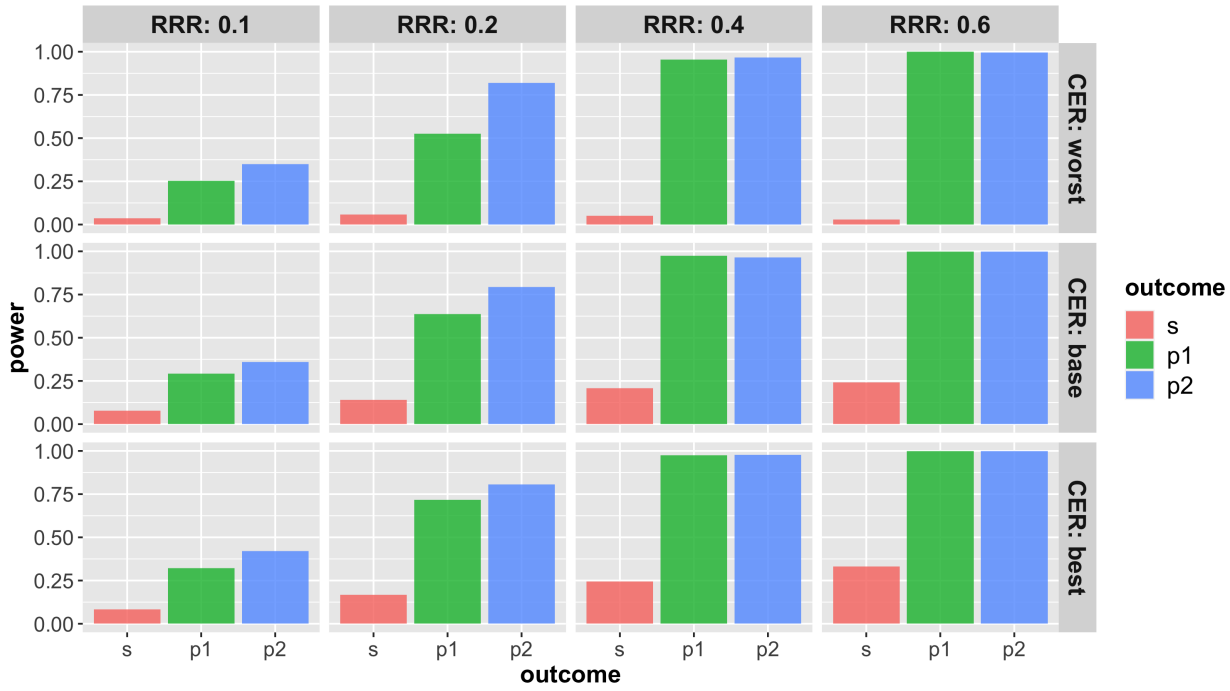


Figure 3. Under the true RRR=40% and 60%, power for outcomes p1 and p2 is very high but decreases to as the RRR becomes less than 20%. Power for outcome s is low for all scenarios, reaching roughly 25% for a true RRR=60% and the best case CER.

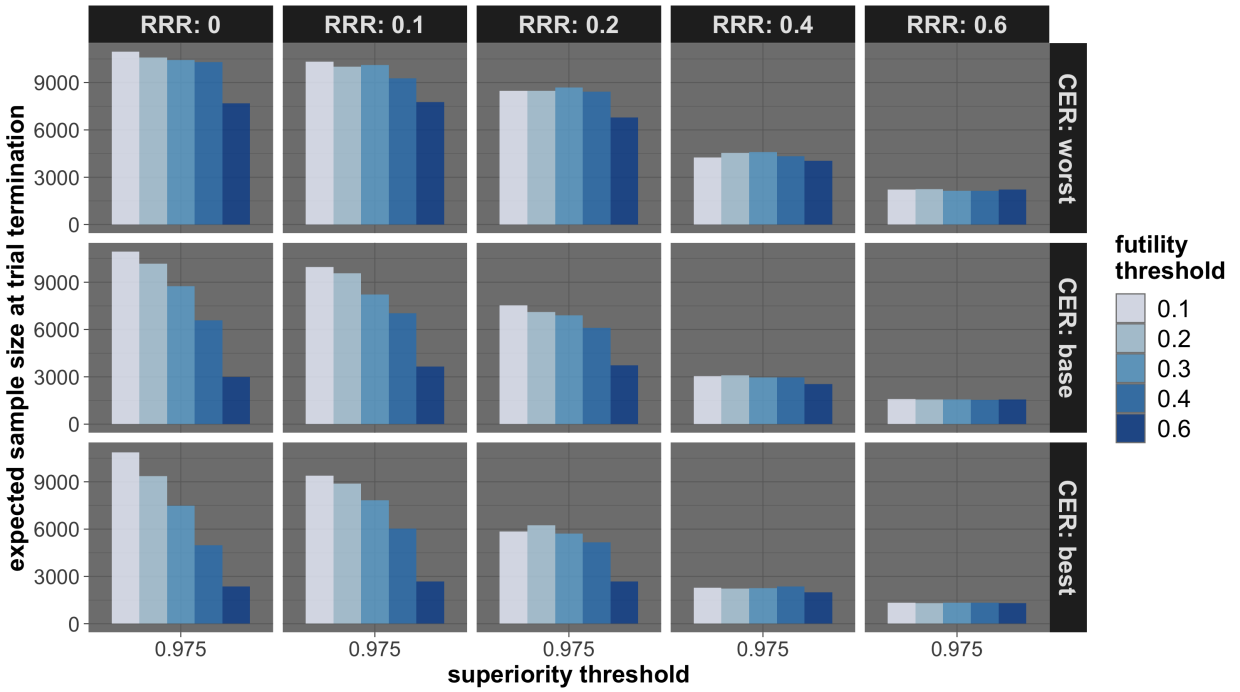
Figure 3: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels or legend). Observed power for each outcome is presented by colored bars. Power estimates by control event rates are presented by the rows and correspond to a superiority threshold of 0.975.



Expected Sample Size

Figure 4 presents the expected (mean) sample size at trial termination. For true RRR=0, 10% and 20%, expected sample sizes were consistently high, albeit with some notable reductions associated with use of a RRR=40% futility stopping threshold. For true of RRR=40% and RRR=60%, expected sample sizes were lower and decreased as the CER improved (worst to best).

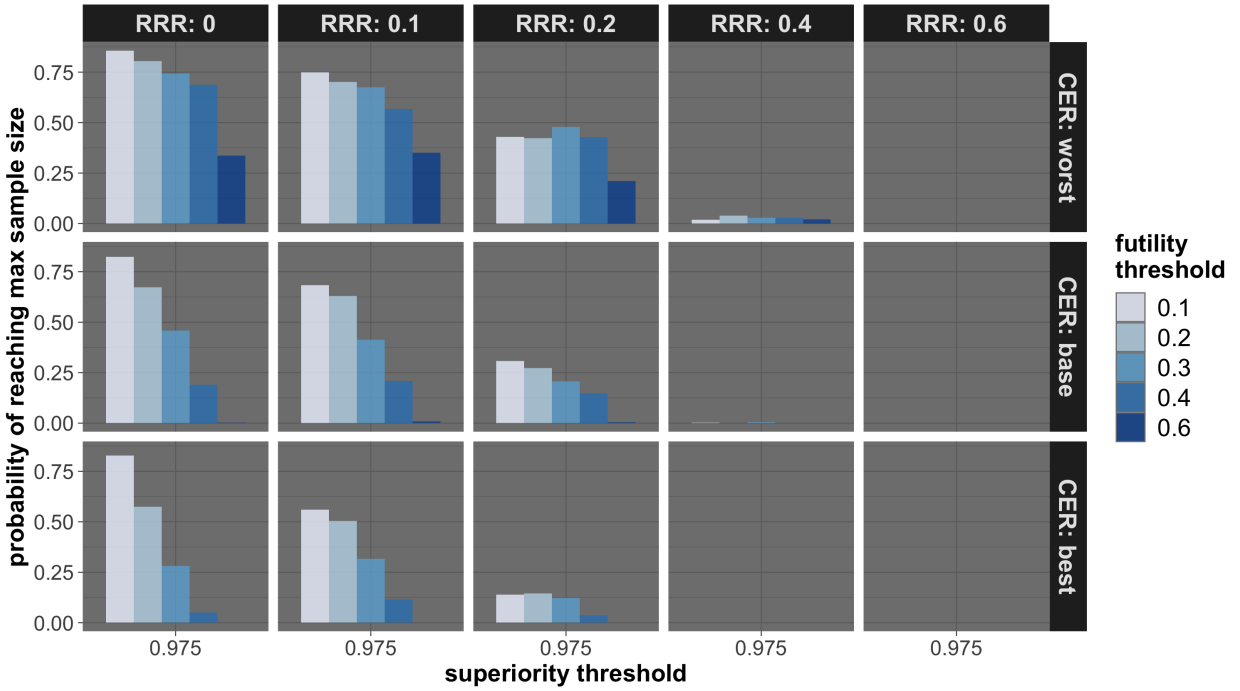
Figure 4: Expected sample size at trial termination. Results by control event scenarios are presented by rows. Results by relative risk reductions are presented by columns. Results by futility thresholds are presented by the color of the bars (see legend).



Probability of Reaching Maximum Sample Size

Figure 5 shows the probability of reaching the maximum allowed sample size of the trial. This probability was moderate to high for the true RRR=0, 10% and 20%, and generally decreased as the CER improved. For RRR=40% and RRR=60%, this probability was negligible for all cases.

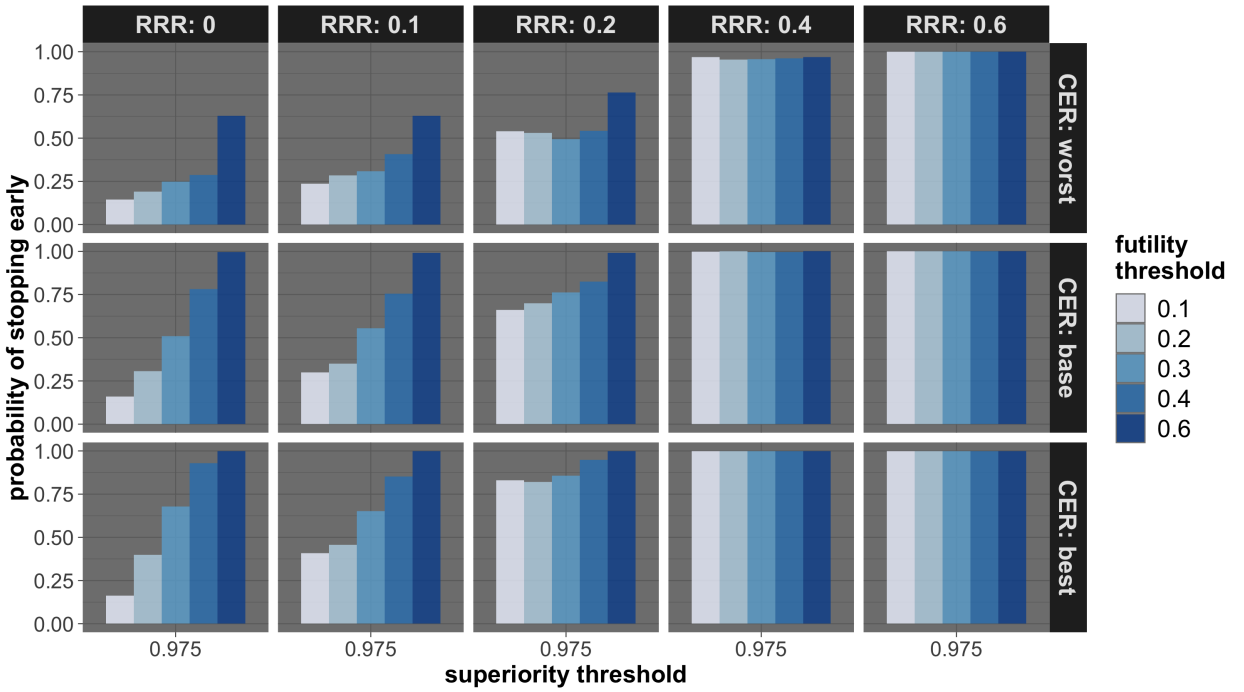
Figure 5: Probability of reaching maximum sample size for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).



Probability of Stopping Early

The probability of stopping early was obtained overall, for superiority and for futility. Figure 6 displays the overall probability of stopping early, Figure 7(a) displays the probability of stopping early for futility and figure 7(b) displays the probability of stopping early for superiority. The overall probability of stopping early is strongly positively correlated with the CER and RRR. When the simulated RRR=0%, 10% and 20%, the overall probability of stopping early is also highly correlated to the futility threshold.

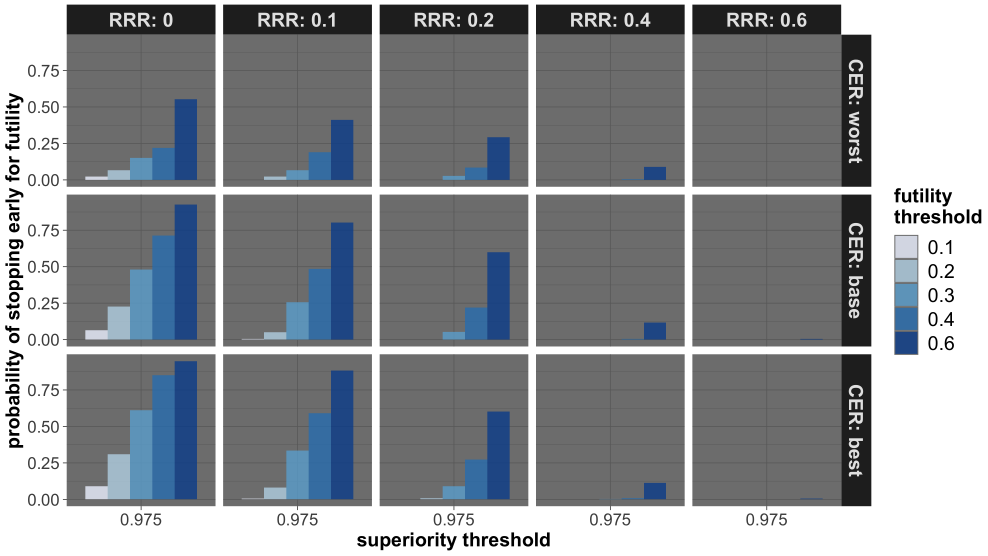
Figure 6: Probability of stopping early due to futility or superiority for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), three superiority thresholds (x axis) and three futility thresholds (legend).



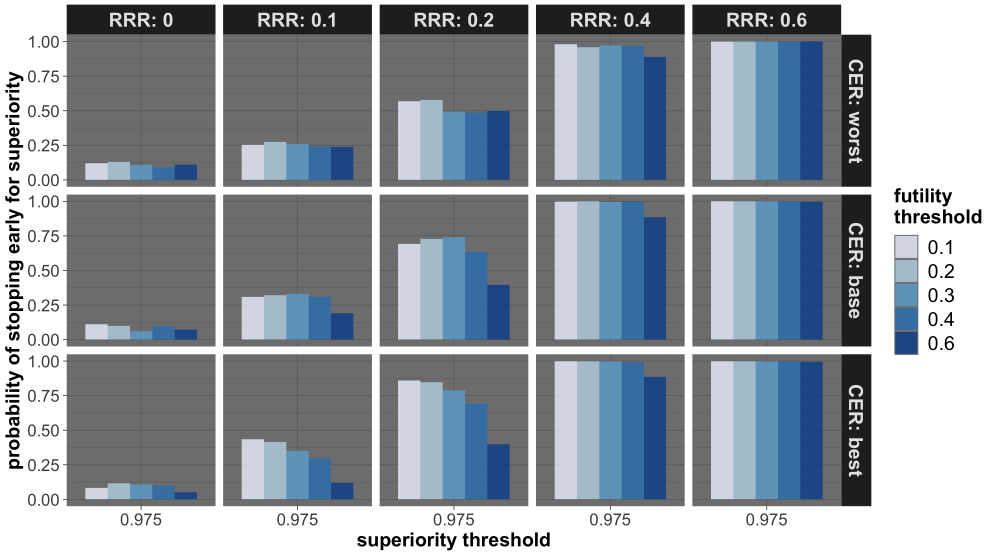
The RRR=40% futility threshold results in consistently greater than 80-90% probability of stopping early for the base and best-case CER scenarios when true RRR=0% (similar trend observed for expected sample size results Figure 3). Stopping early for futility when the simulated RRR=0% is likely with futility thresholds of >30%, but less so with a futility threshold of 20%. A futility threshold of >40% also results in a moderately low probability of stopping early where the true RRR=20%. The probability of stopping early for superiority is close to 100% for all CERs when the true RRR=40% or RRR=60%, regardless of futility threshold. For the true RRR=20%, there is moderate probability of stopping early which increases with improving CER and decreases with higher futility thresholds. The lower probability of stopping for superiority when the true RRR=0 explained by the high probability of stopping for futility in this setting.

Figure 7: Probability of stopping early due to futility, and stopping early due to superiority. Stopping probabilities are presented for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).

(a)



(b)

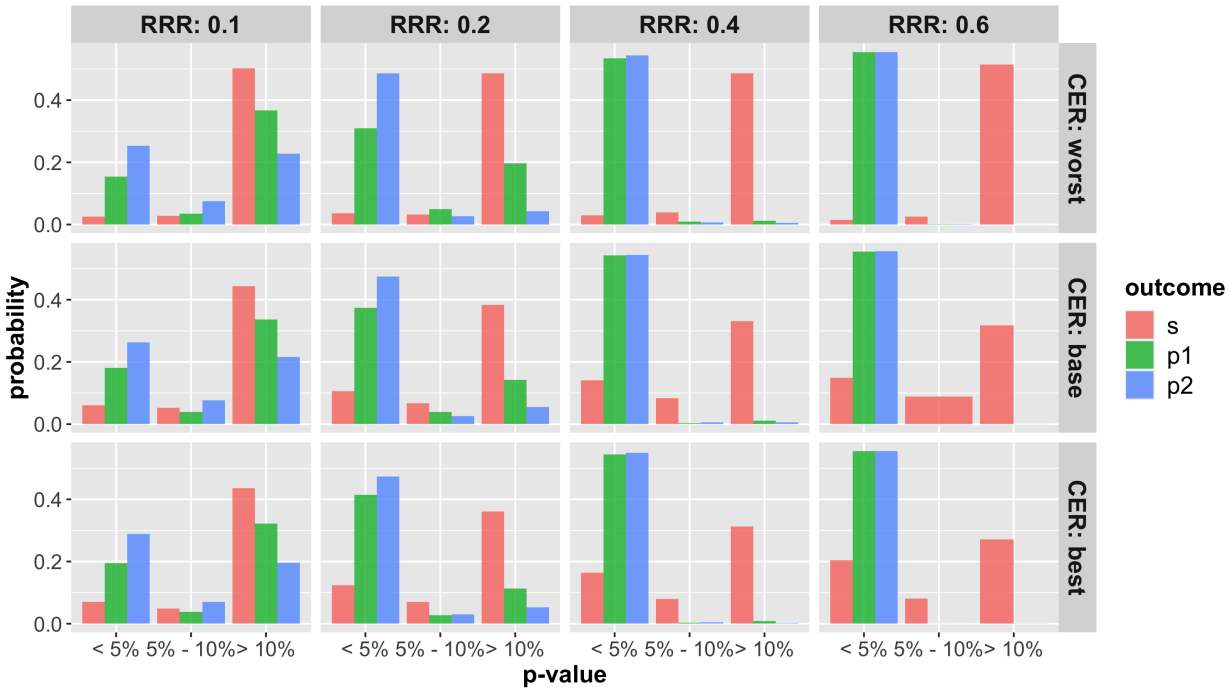


P-values at trial termination when a true effect exists

Figure 8 presents the categorical distribution of p-values (<5%, 5-10%, or >10%) upon trial termination (stopping early or reaching max. allowed sample size). Figure 9(a) and 9(b) presents the divide of p-values when stopping for futility (a) and superiority (b), respectively.

The overall probability of observing a p-value less than 5% (i.e., a conventionally statistically significant difference) is high or close to 100% in scenarios when RRR=40% or 60% across all CER scenarios for outcomes p1 and p2. Outcome s has a greater proportion of p-values being >10% in these scenarios.

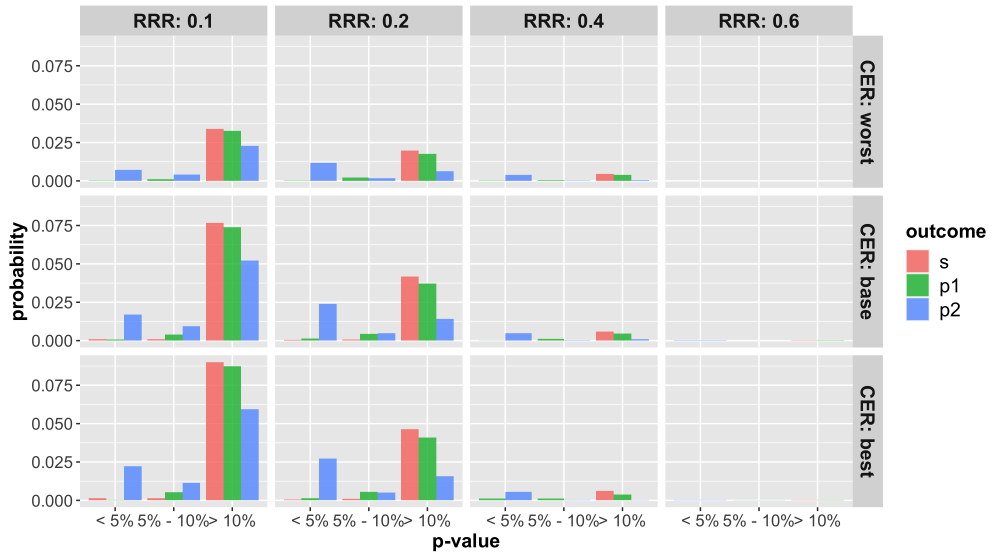
Figure 8: Overall probability at trial termination that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10%. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios.



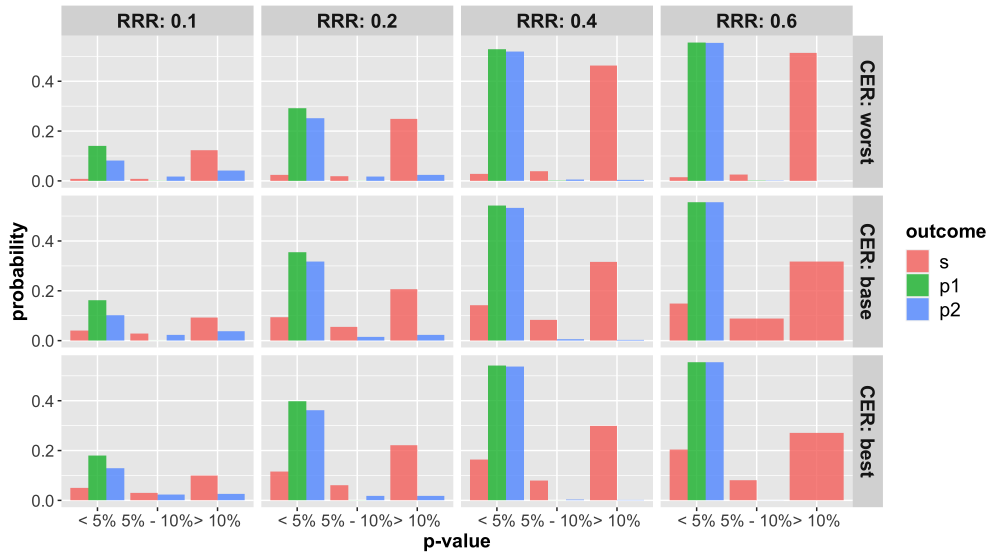
In the situations where the trial is stopped early for futility, the small counts for RRR=40% and 60% can be explained by the low probability of stopping early for futility under these scenarios (Figure 7(a)). When the trial is stopped for superiority when the true RRR=40% or 60%, close to 100% of all p-values for outcomes p1 and p2 are smaller than 5%. Under all RRR and CER scenarios, the majority of p-values for outcomes s remain greater than 10%.

Figure 9: Probability that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10% for cases where trial was (a) stopped for futility; (b) stopped for superiority. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios. Note: the denominator in each figure is the number of simulations (not the number of trials stopped for futility (a) or superiority (b), and thus, the proportions do not add up to 100% within one figure. Further, (a) and (b) do not include simulations where the trial went to the max. allowed sample size. The bars should be interpreted with respect to the relative proportion that fit in each category.

(a)



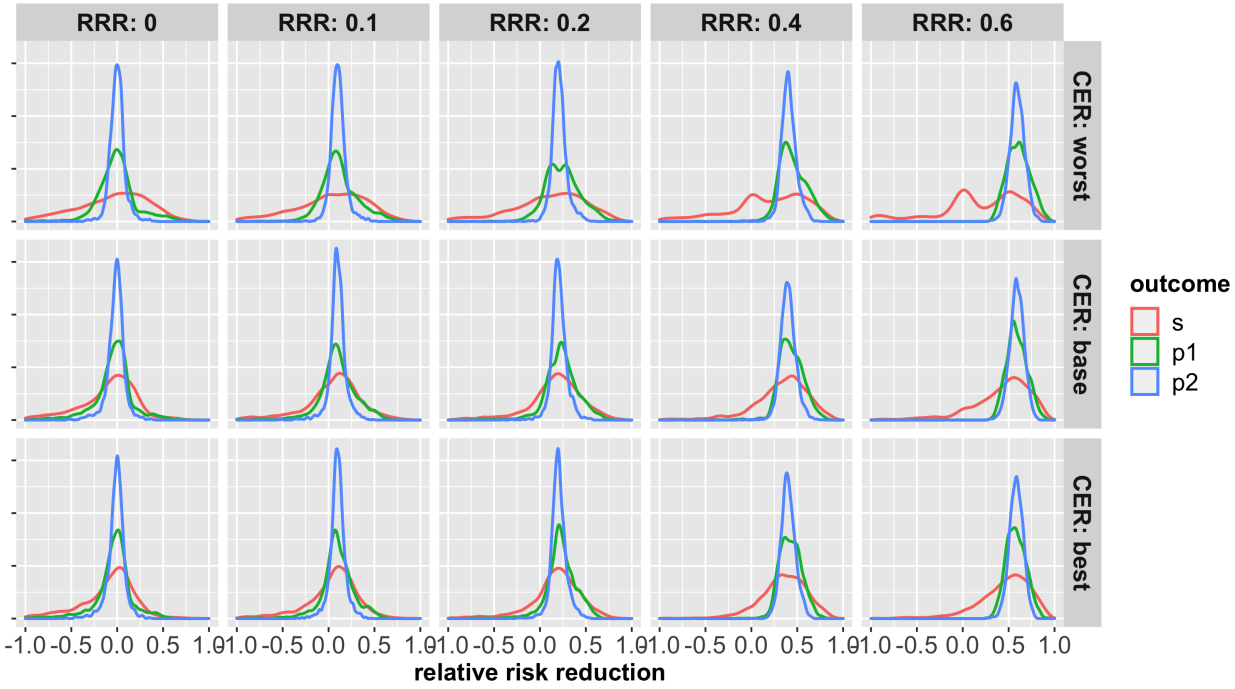
(b)



Relative risk reduction estimates at trial termination

Figure 10 presents the distribution of relative risk reduction estimates upon trial termination. Figure 11(a) and 11(b) present the distribution of relative risk reduction estimates from trials stopped early for futility and superiority, respectively. As expected, the estimates of p1 and p2 exhibit much larger precision.

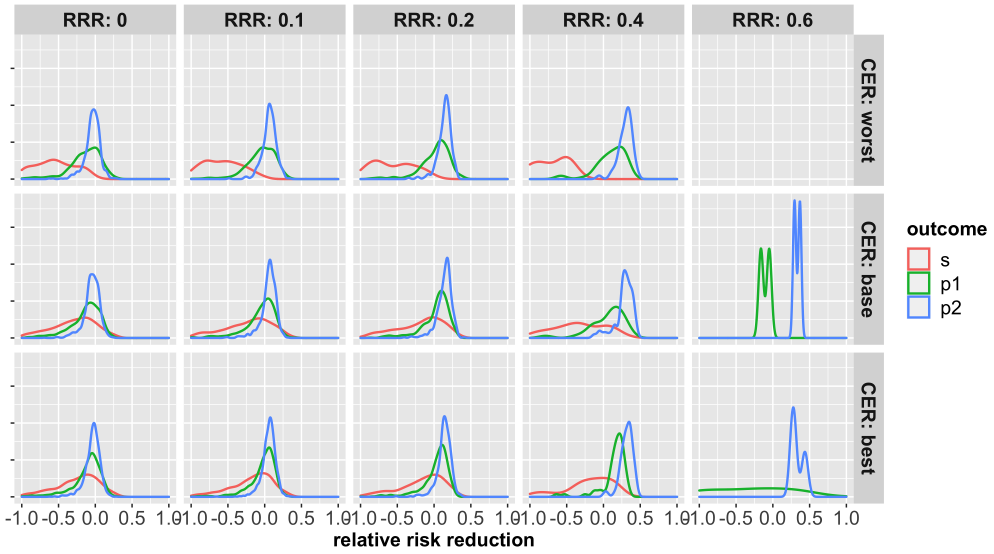
Figure 10: Distribution of relative risk reduction estimates (smoothed by a kernel density estimator) for the three control event rates (CER – rows), four relative risk reductions (RRR – columns) and the three outcomes (legend).



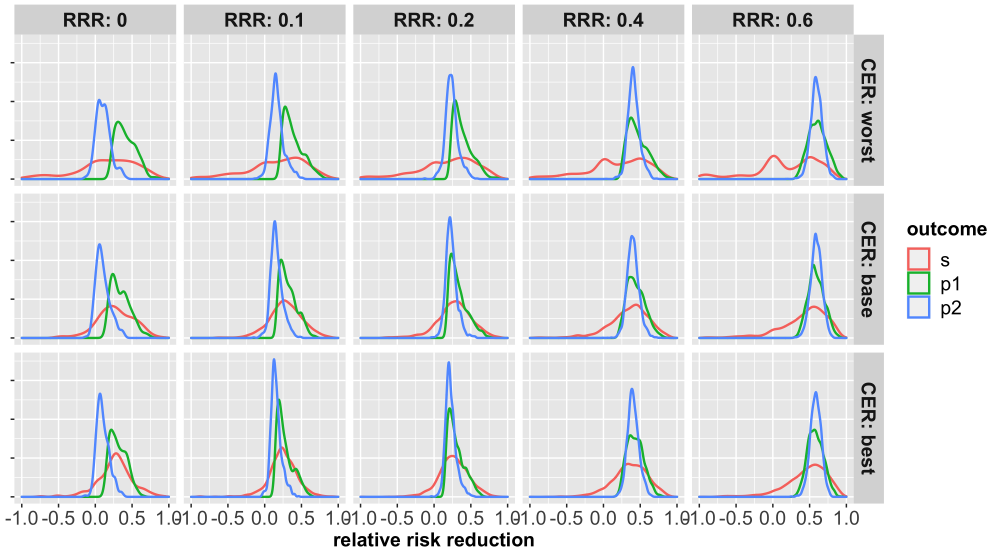
When stopping for futility, outcome s incurs small to moderate downward bias for all scenarios. For true RRR=0 and 20% and across all CER scenarios, both outcomes p1 and p2 show negligible bias, though p2 shows greater precision. When stopping for futility, the estimates for RRR = 40% and 60% are unstable, but the kernel density smoother obscures this. When stopping for superiority, outcome s is very diffuse in the worst case CER scenario, but sees its precision improve with the CER. It has marginal upward bias when the true RRR=0 and 20%, but shows little bias for true RRR=40% and 60% for the base and best case CER scenarios. Outcome p2 has greatest precision across all scenarios and has negligible bias. Outcome p1 has moderate to good precision across all scenarios but has moderate upward bias for true RRR=0, 10%, and 20%.

Figure 11: Distribution of relative risk reduction estimates after stopping early for (a) futility; (b) superiority. Results are presented for the three control event rates by rows, four relative risk reductions (by columns) and the three outcomes (legend).

(a)



(b)



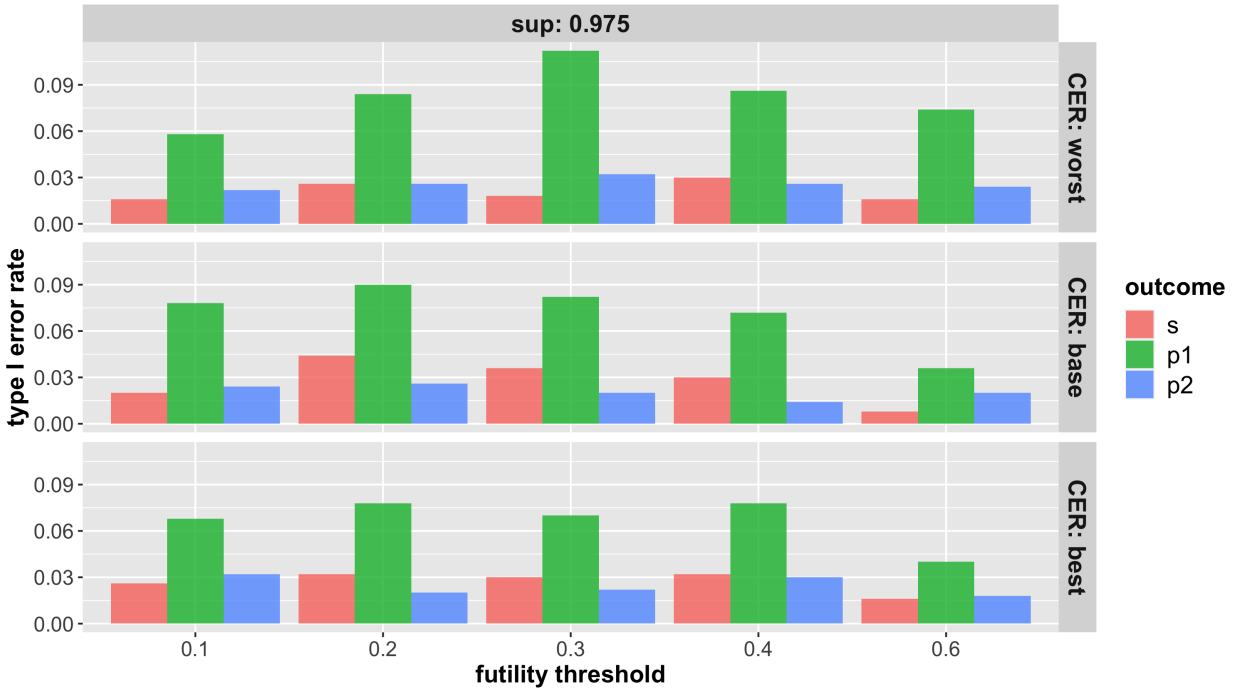
Stopping with moderately permissive outcome (p1)

Maximum allowed sample size of 12,000 with interim analyses conducted every 2,000 patients

Both superiority and futility decisions were made with respect to the moderately permissive outcome (p1). Unless stated otherwise, the futility bound used was 0.4 and a probability threshold of 0.99 was required to stop for futility. The number of expected interim analyses can be thought of as the expected number of patients at trial determination divided by the batch size, which here is 2,000 patients. The plots below can be interpretated in the same manner as those above.

Type I error (false positive risk)

Figure 12: Overall type I error rate based on moderately permissive (p1) outcome based stopping rules.Observed type one error rate for each outcome is presented by colored bars (see legend). Results by the three categories of control event rates are by rows. Results by the futility thresholds are presented on the x-axis.



Power

Figure 13: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels). Power estimates by control event rates are presented by the rows. Power estimates by the futility thresholds are presented by the bar color (see legend).

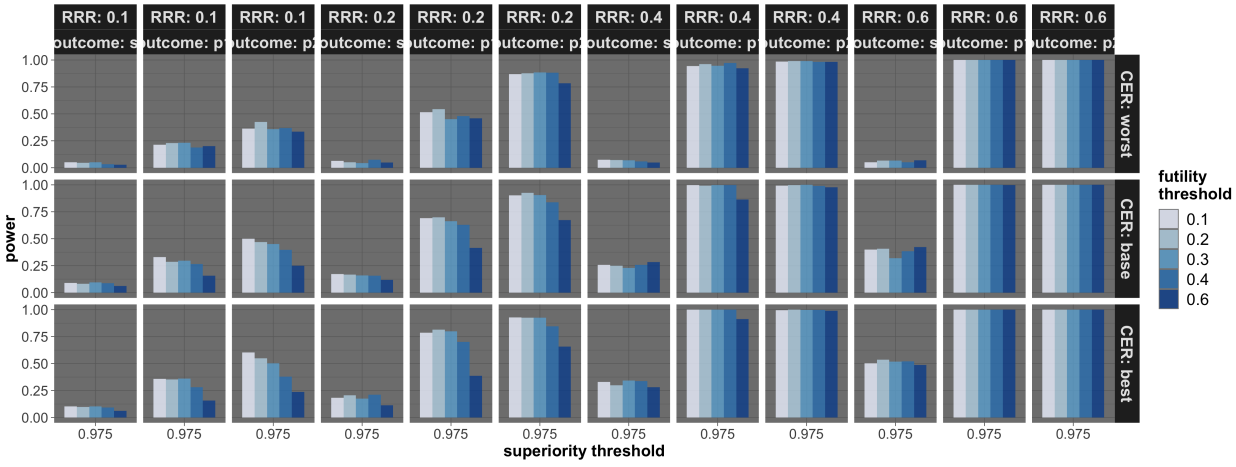
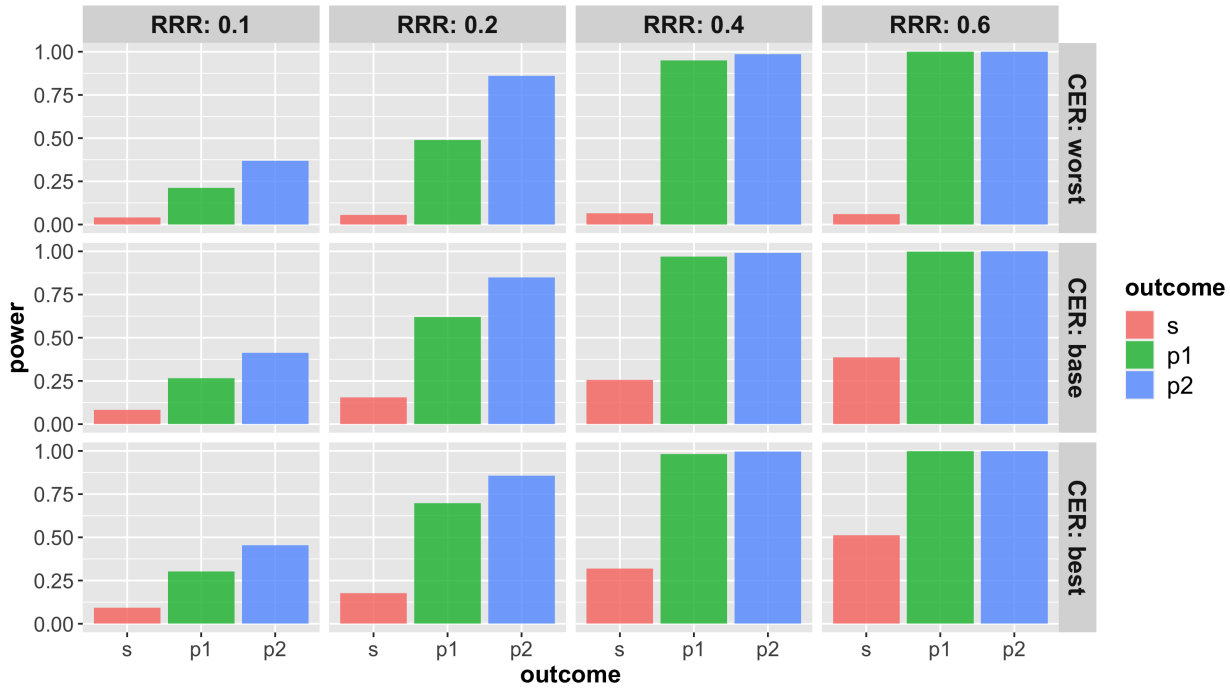
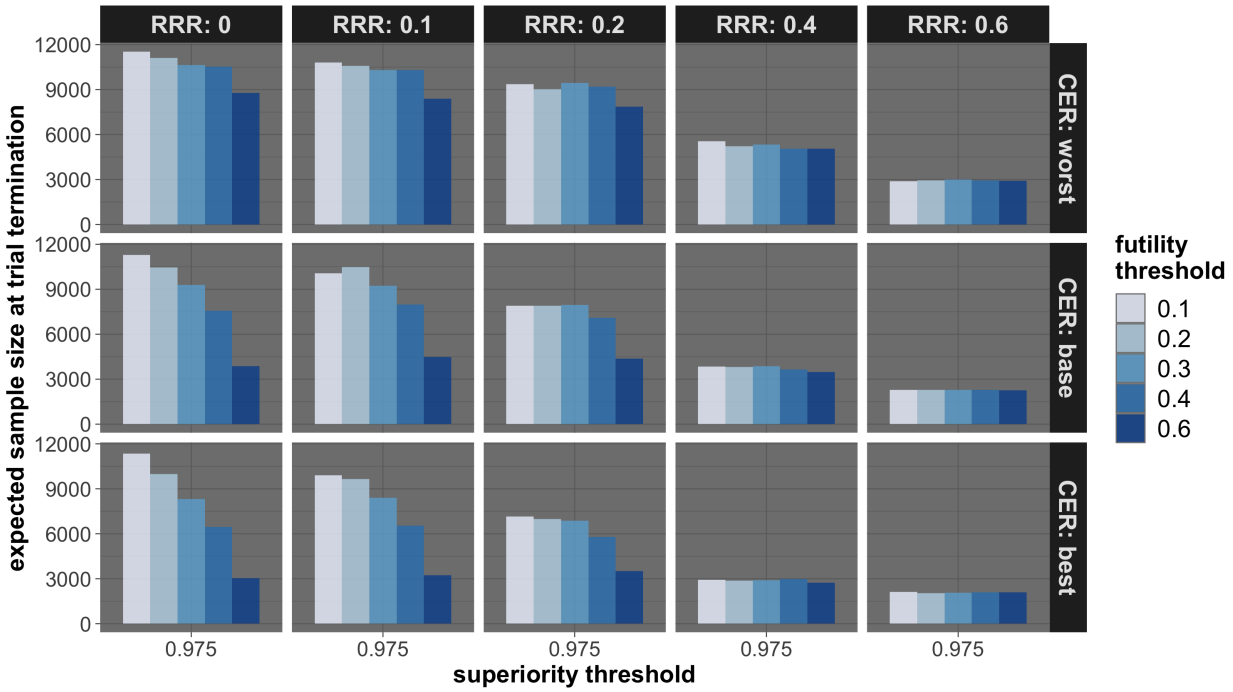


Figure 14: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels or legend). Observed power for each outcome is presented by colored bars. Power estimates by control event rates are presented by the rows and correspond to a superiority threshold of 0.975.



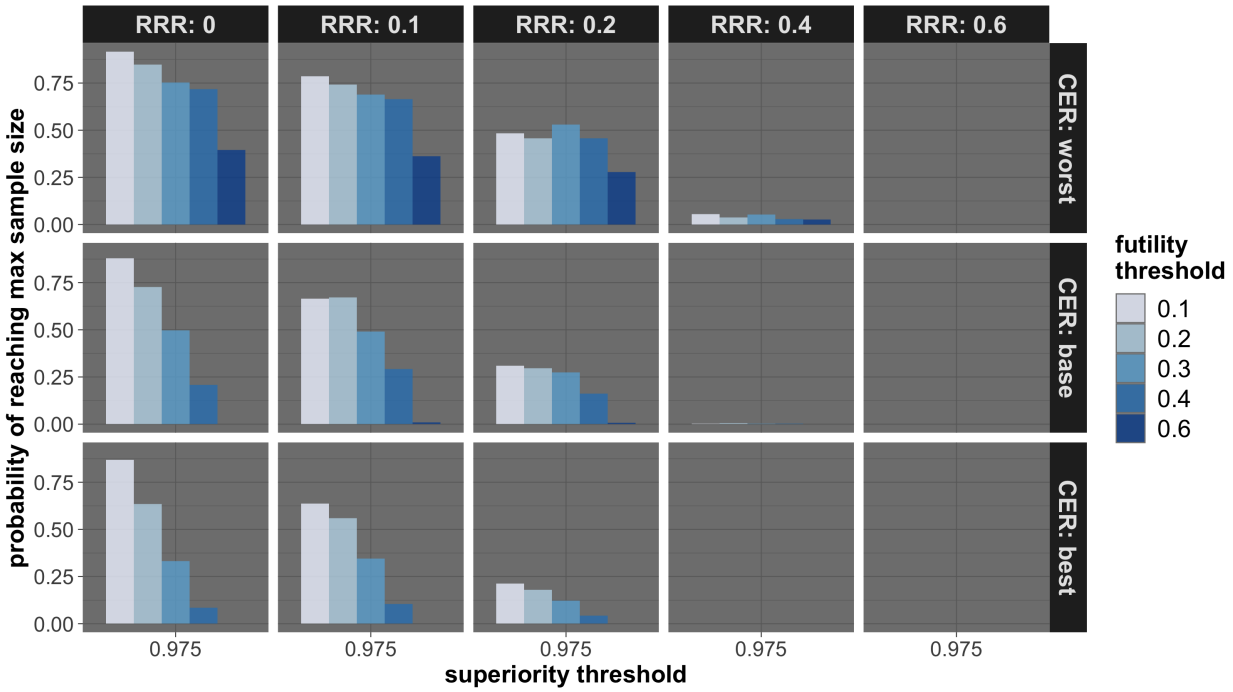
Expected Sample Size

Figure 15: Expected sample size at trial termination. Results by control event scenarios are presented by rows. Results by relative risk reductions are presented by columns. Results by futility thresholds are presented by the color of the bars (see legend).



Probability of Reaching Maximum Sample Size

Figure 16: Probability of reaching maximum sample size for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).



Probability of Stopping Early

Figure 17: Probability of stopping early due to futility or superiority for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), three superiority thresholds (x axis) and three futility thresholds (legend).

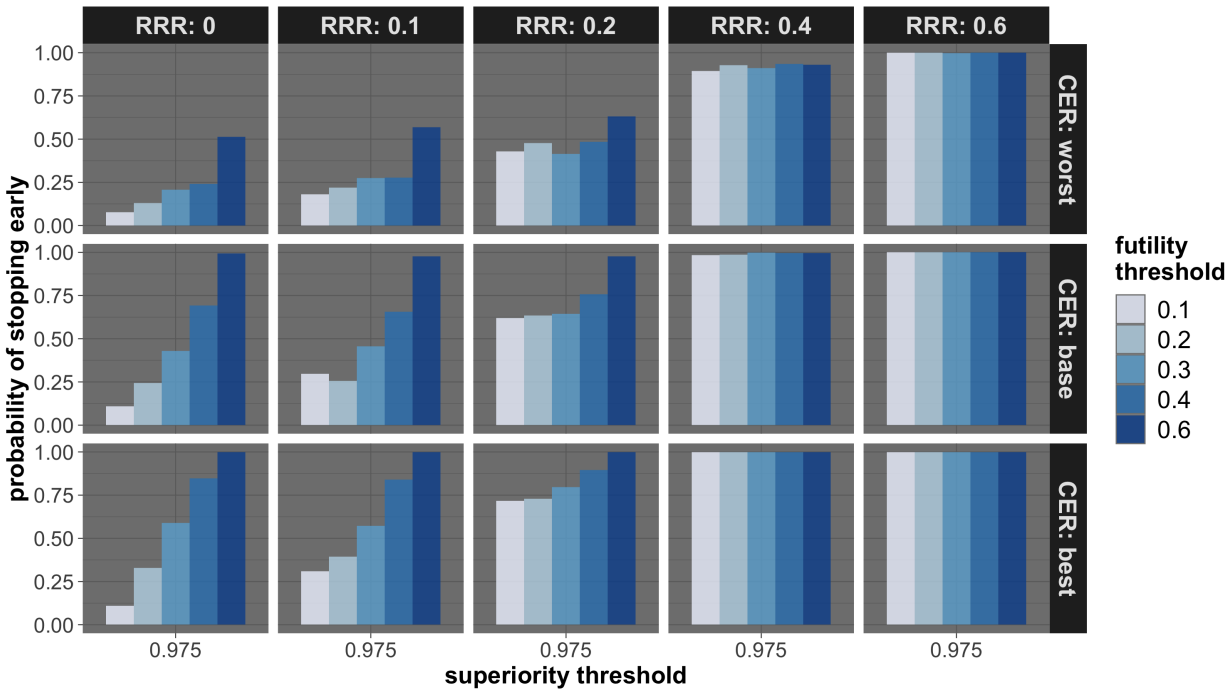
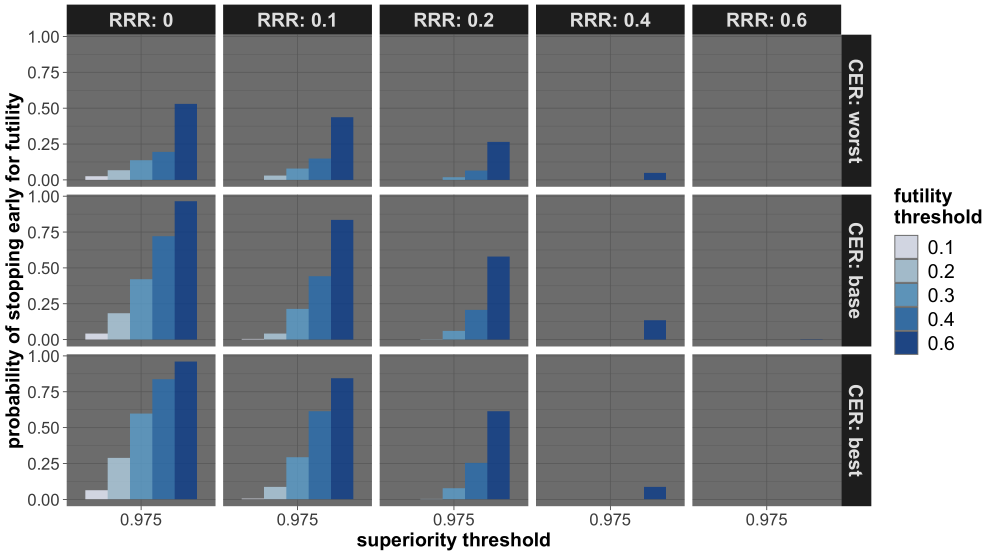
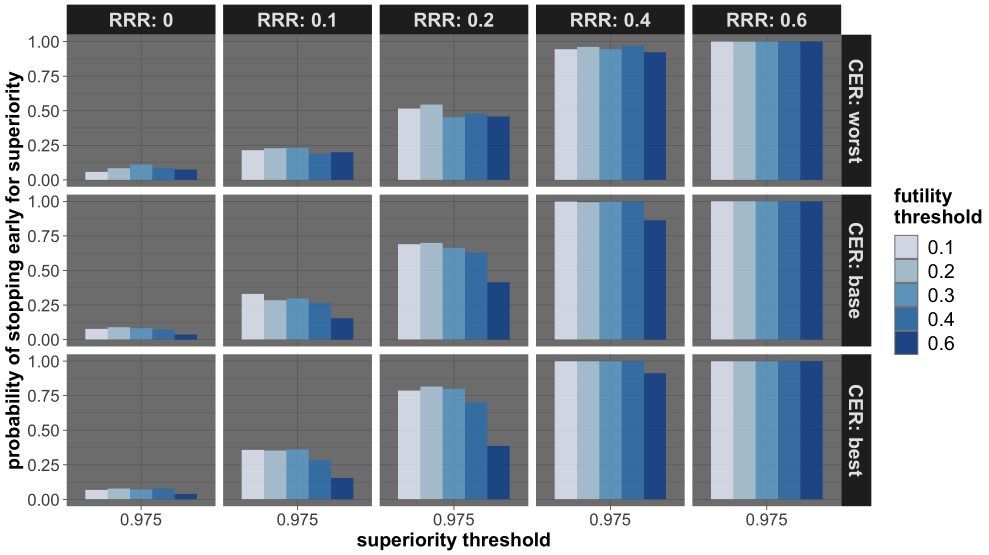


Figure 18: Probability of stopping early due to futility, and stopping early due to superiority. Stopping probabilities are presented for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).

(a)



(b)



P-values at trial termination when a true effect exists

Figure 19: Overall probability at trial termination that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10%. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios.

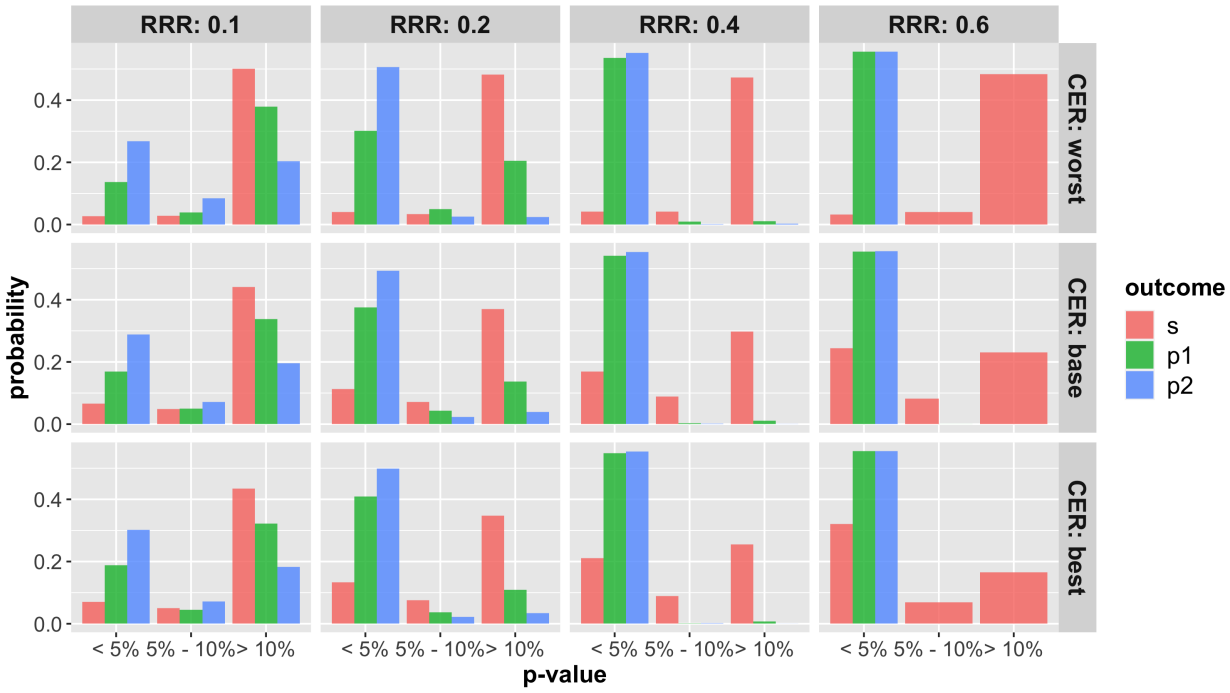
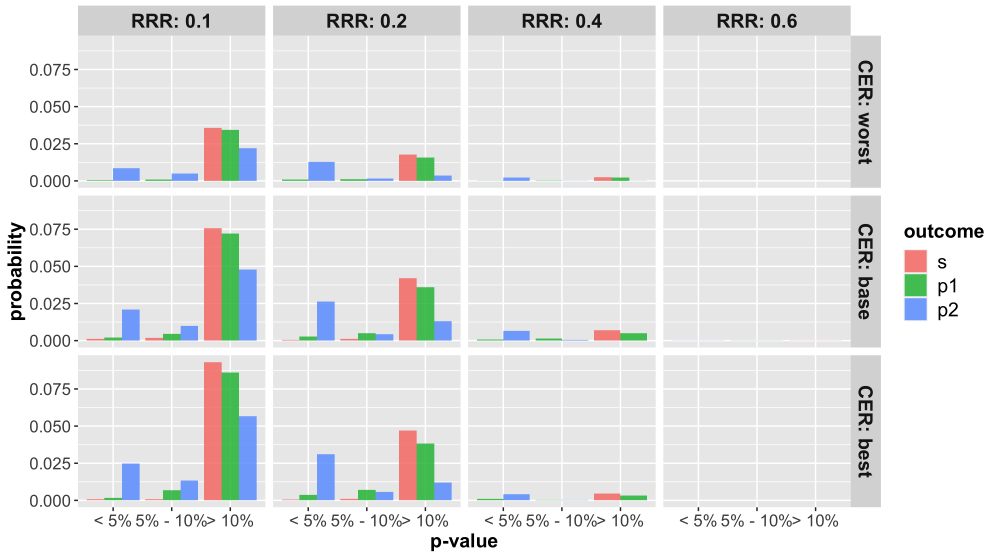
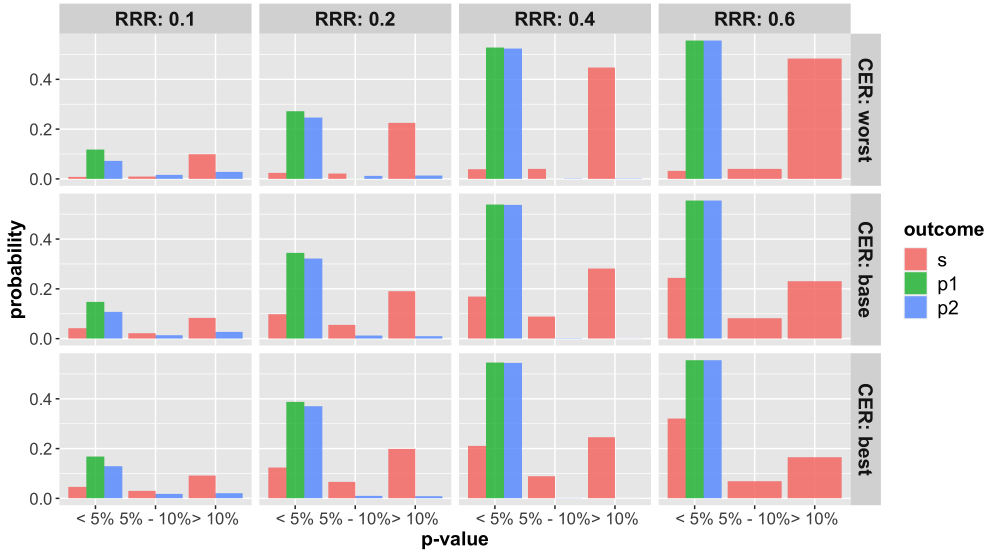


Figure 20: Probability that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10% for cases where trial was (a) stopped for futility; (b) stopped for superiority. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios. Note: the denominator in each figure is the number of simulations (not the number of trials stopped for futility (a) or superiority (b), and thus, the proportions do not add up to 100% within one figure. Further, (a) and (b) do not include simulations where the trial went to the max. allowed sample size. The bars should be interpreted with respect to the relative proportion that fit in each category.

(a)



(b)



Relative risk reduction estimates at trial termination

Figure 21: Distribution of relative risk reduction estimates (smoothed by a kernel density estimator) for the three control event rates (CER – rows), four relative risk reductions (RRR – columns) and the three outcomes (legend).

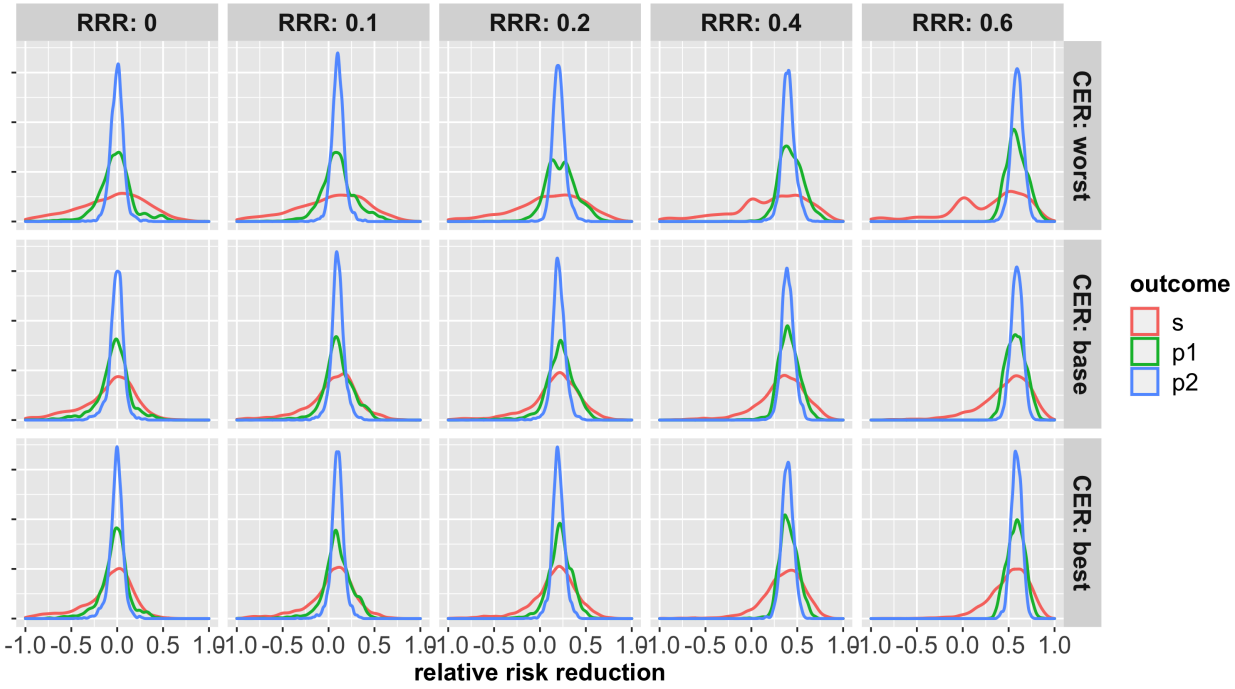
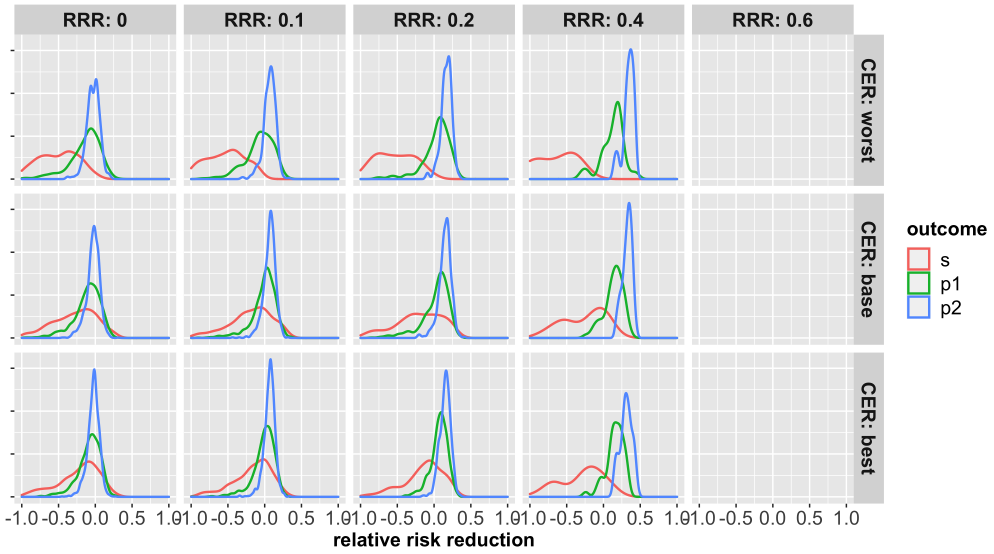
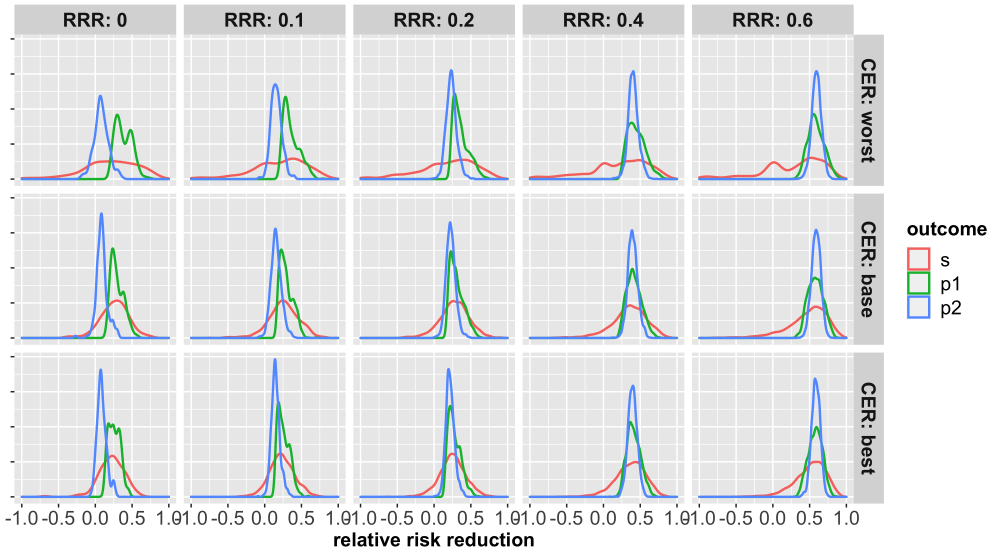


Figure 22: Distribution of relative risk reduction estimates after stopping early for (a) futility; (b) superiority. Results are presented for the three control event rates by rows, four relative risk reductions (by columns) and the three outcomes (legend).

(a)



(b)



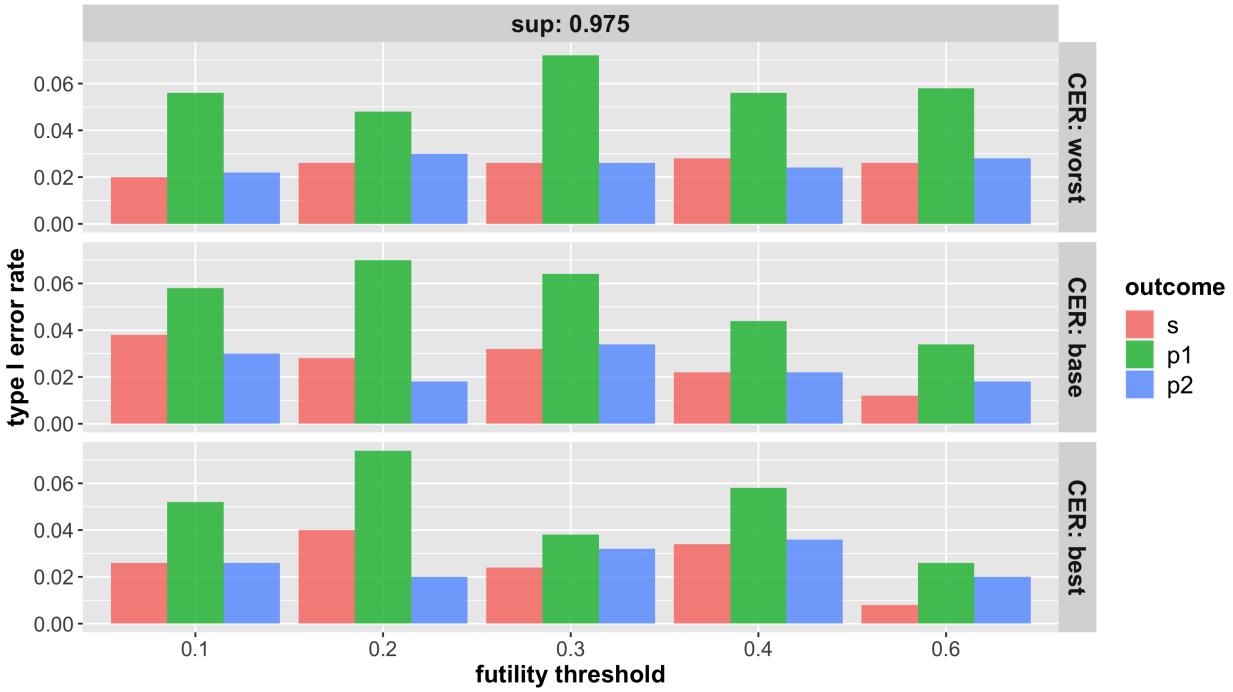
Stopping with moderately permissive outcome (p1)

Maximum allowed sample size of 12,000 with interim analyses conducted every 3,000 patients

Both superiority and futility decisions were made with respect to the moderately permissive outcome (p1). Unless stated otherwise, the futility bound used was 0.4 and a probability threshold of 0.99 was required to stop for futility. The number of expected interim analyses can be thought of as the expected number of patients at trial determination divided by the batch size, which here is 3,000 patients. The plots below can be interpretated in the same manner as those above.

Type I error (false positive risk)

Figure 23: Overall type I error rate based on moderately permissive (p1) outcome based stopping rules.Observed type one error rate for each outcome is presented by colored bars (see legend). Results by the three categories of control event rates are by rows. Results by the futility thresholds are presented on the x-axis.



Power

Figure 24: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels). Power estimates by control event rates are presented by the rows. Power estimates by the futility thresholds are presented by the bar color (see legend).

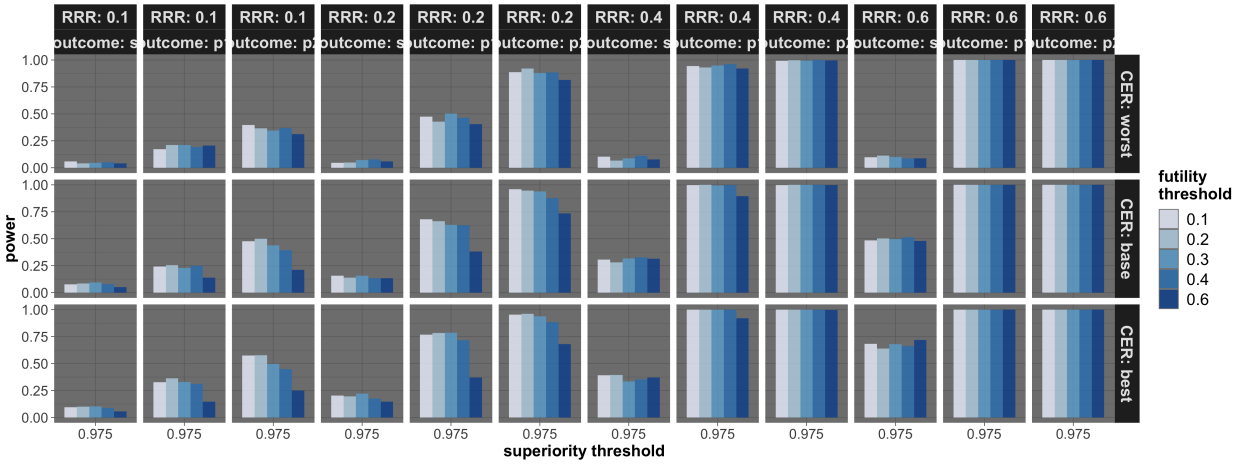
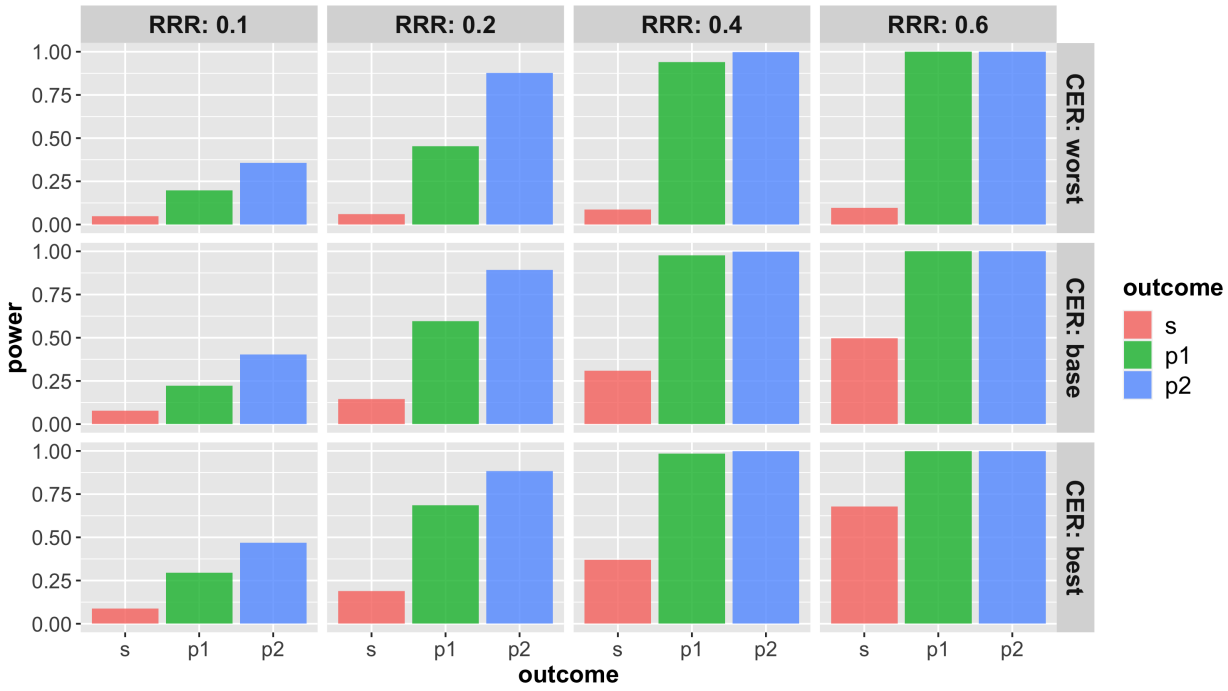
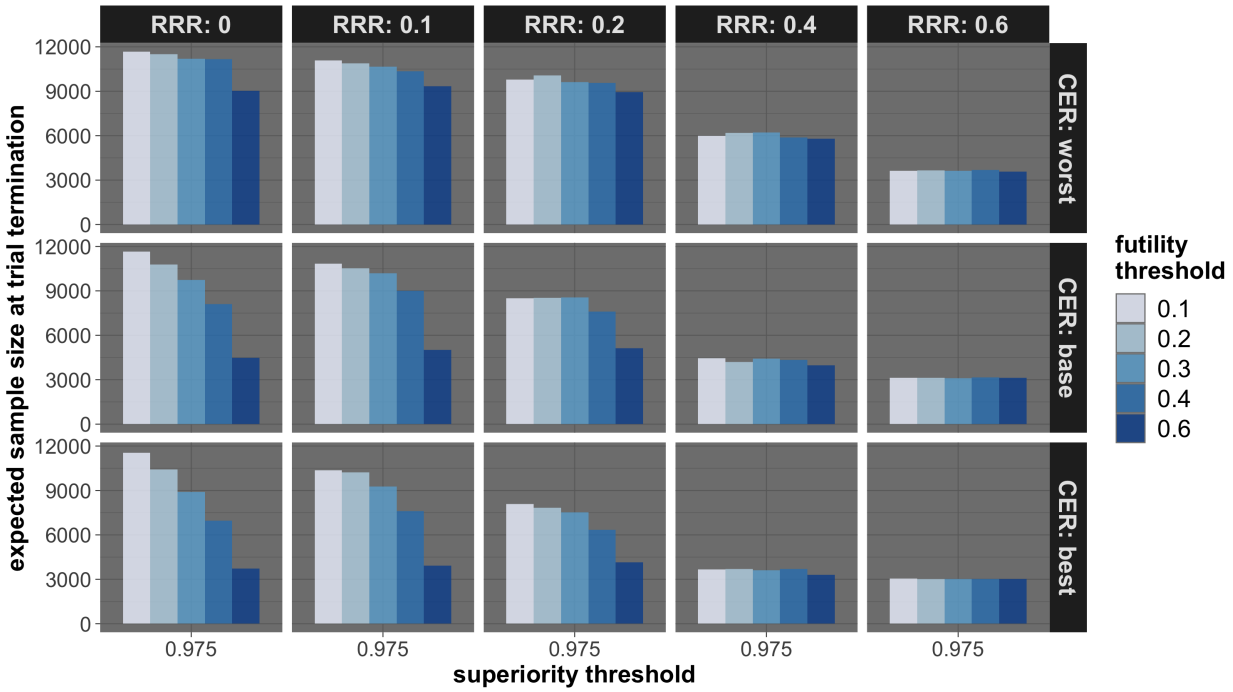


Figure 25: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels or legend). Observed power for each outcome is presented by colored bars. Power estimates by control event rates are presented by the rows and correspond to a superiority threshold of 0.975.



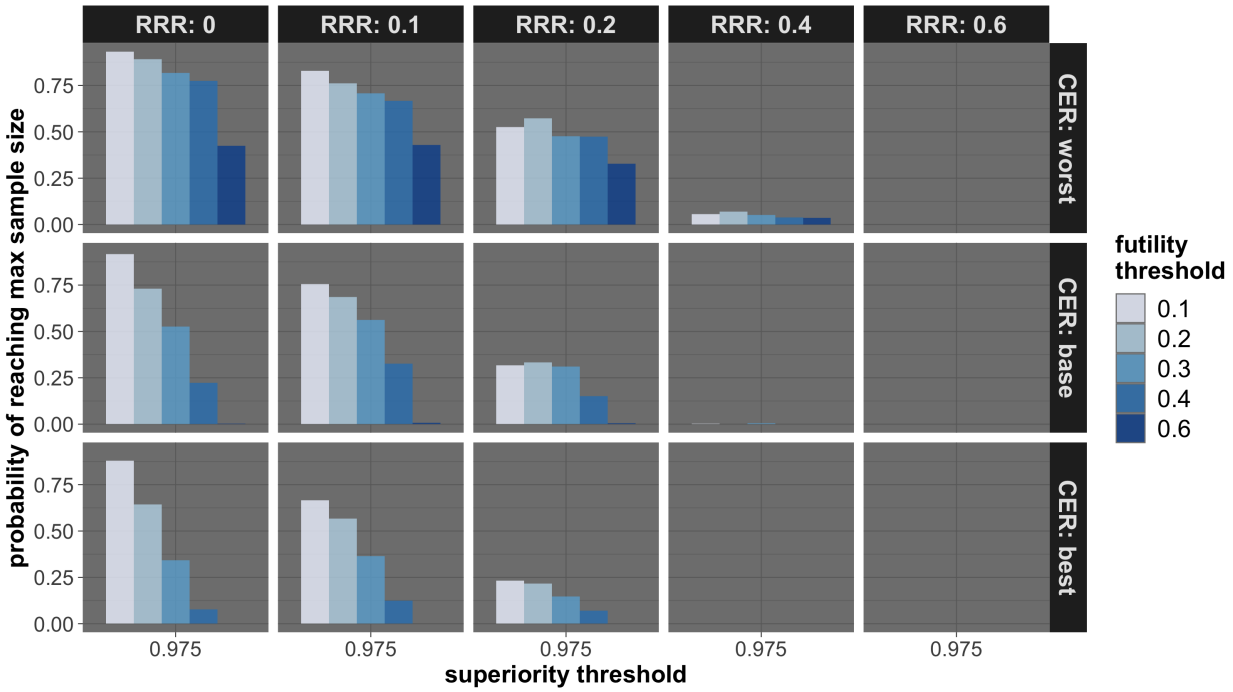
Expected Sample Size

Figure 26: Expected sample size at trial termination. Results by control event scenarios are presented by rows. Results by relative risk reductions are presented by columns. Results by futility thresholds are presented by the color of the bars (see legend).



Probability of Reaching Maximum Sample Size

Figure 27: Probability of reaching maximum sample size for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).



Probability of Stopping Early

Figure 28: Probability of stopping early due to futility or superiority for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), three superiority thresholds (x axis) and three futility thresholds (legend).

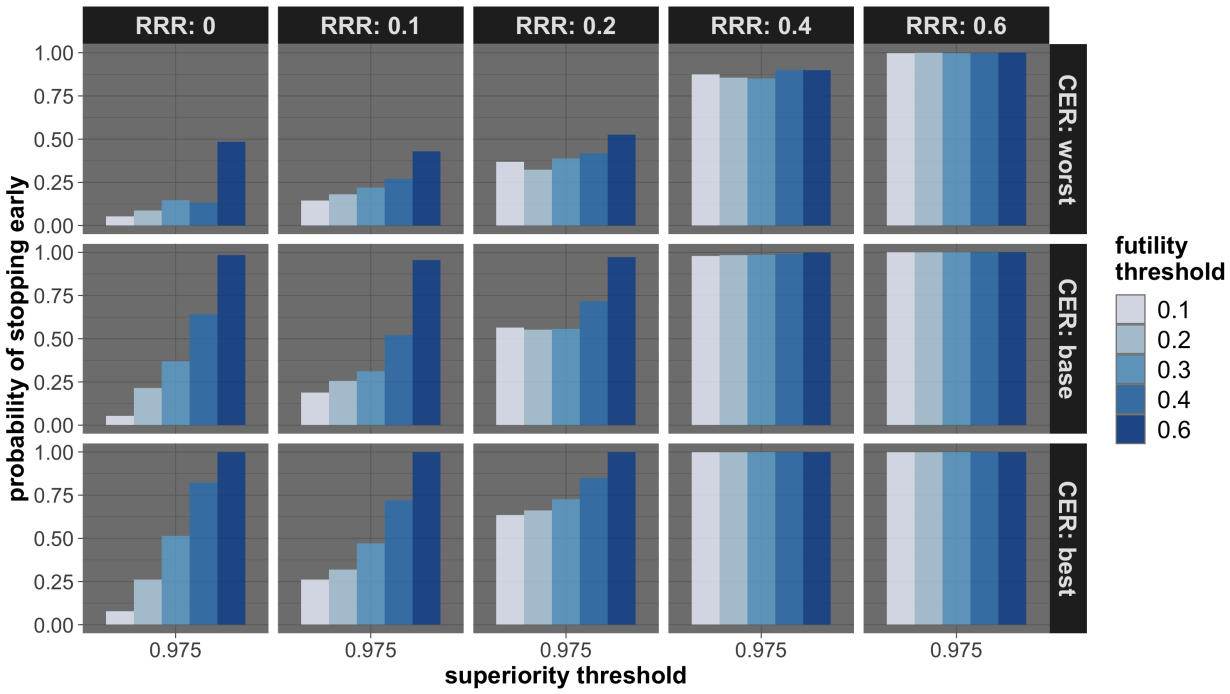
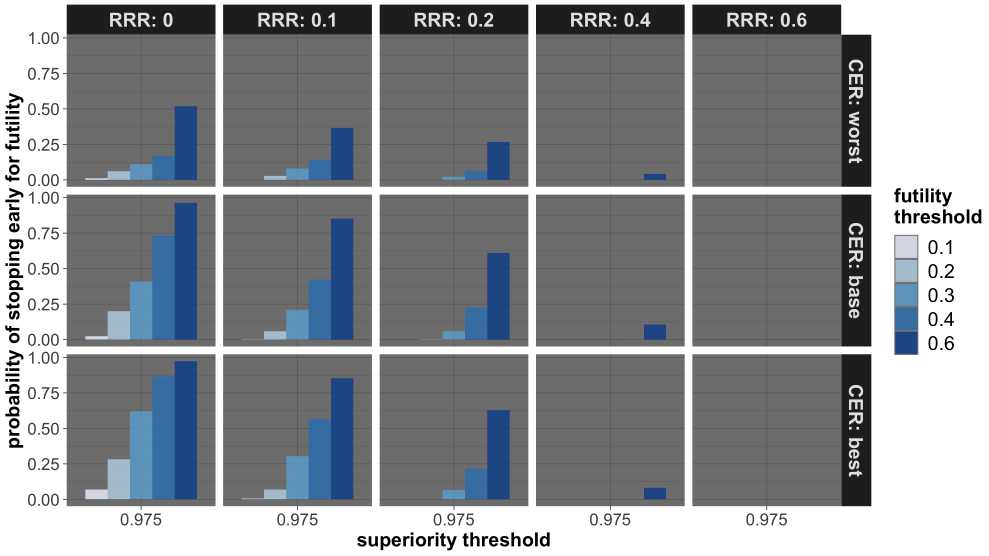
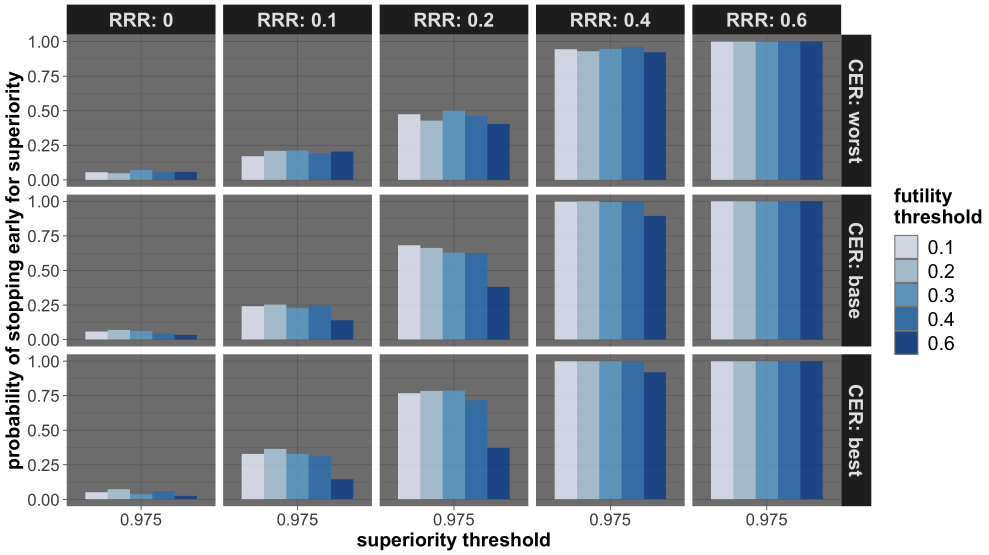


Figure 29: Probability of stopping early due to futility, and stopping early due to superiority. Stopping probabilities are presented for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and three futility thresholds (legend).

(a)



(b)



P-values at trial termination when a true effect exists

Figure 30: Overall probability at trial termination that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10%. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios.

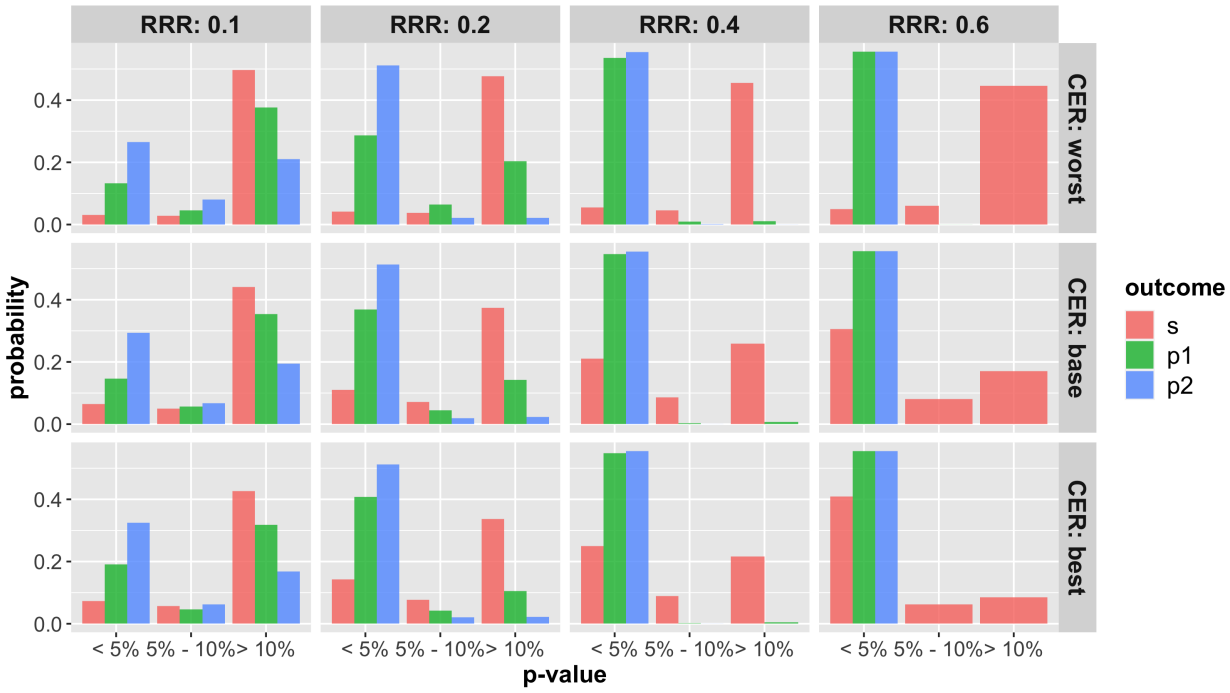
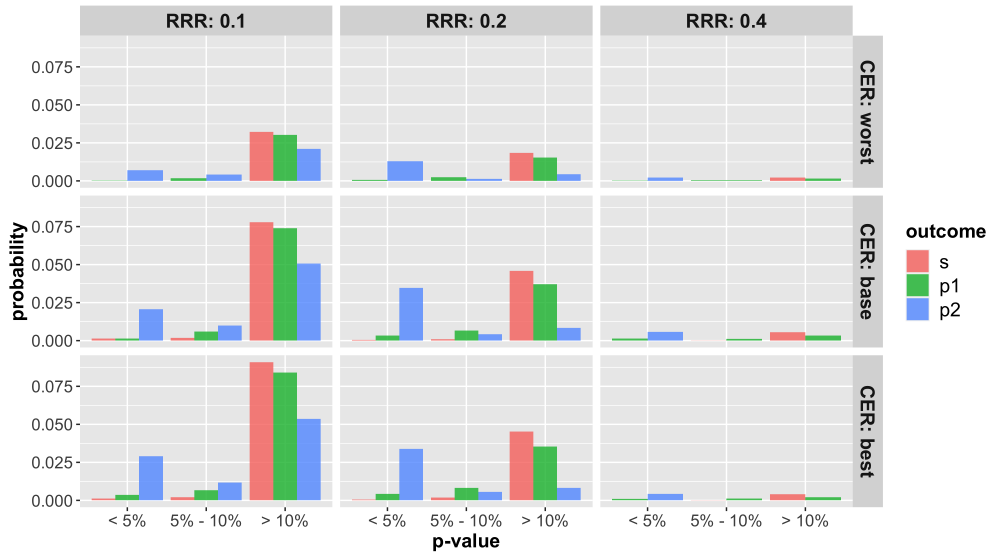
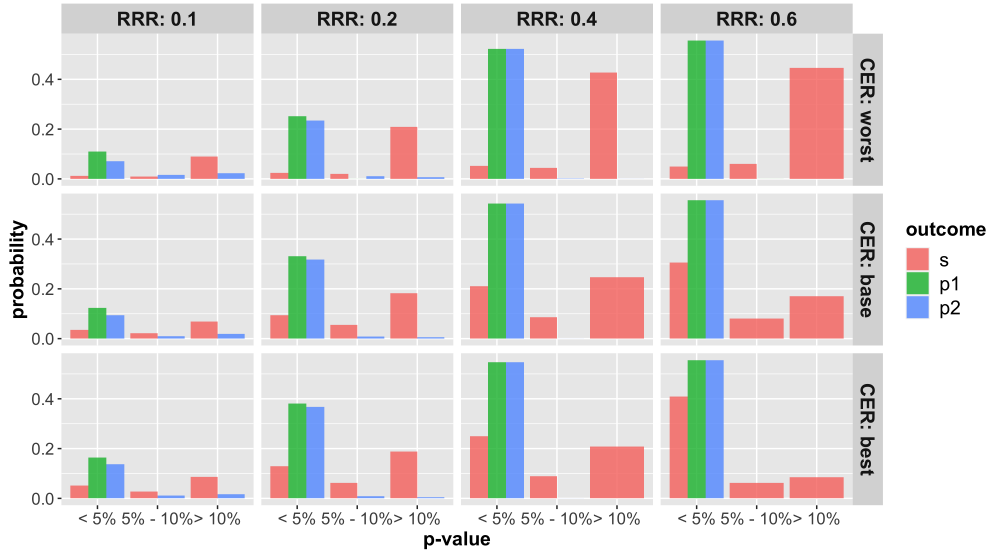


Figure 31: Probability that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10% for cases where trial was (a) stopped for futility; (b) stopped for superiority. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios. Note: the denominator in each figure is the number of simulations (not the number of trials stopped for futility (a) or superiority (b), and thus, the proportions do not add up to 100% within one figure. Further, (a) and (b) do not include simulations where the trial went to the max. allowed sample size. The bars should be interpreted with respect to the relative proportion that fit in each category.

(a)



(b)



Relative risk reduction estimates at trial termination

Figure 32: Distribution of relative risk reduction estimates (smoothed by a kernel density estimator) for the three control event rates (CER – rows), four relative risk reductions (RRR – columns) and the three outcomes (legend).

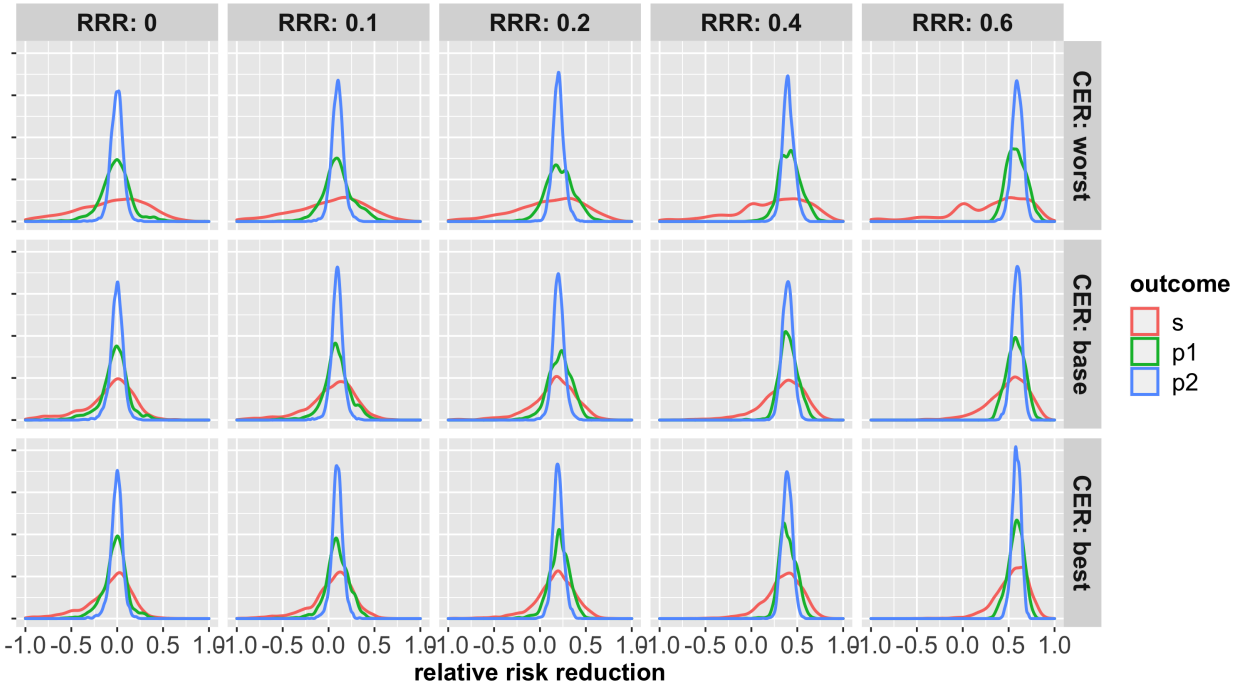
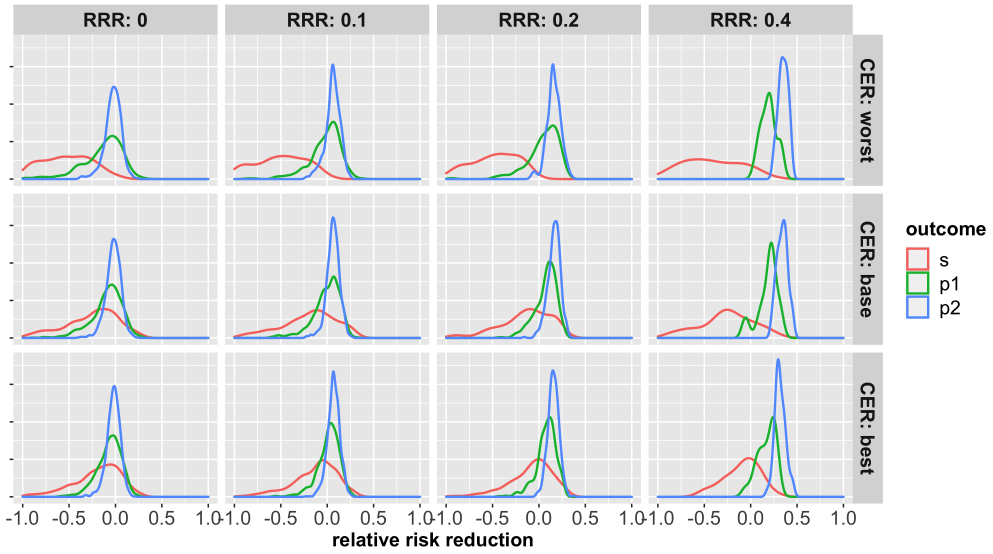


Figure 33: Distribution of relative risk reduction estimates after stopping early for (a) futility; (b) superiority. Results are presented for the three control event rates by rows, four relative risk reductions (by columns) and the three outcomes (legend).

(a)



(b)