Appendix: SEPSIS Trial

2020-10-14

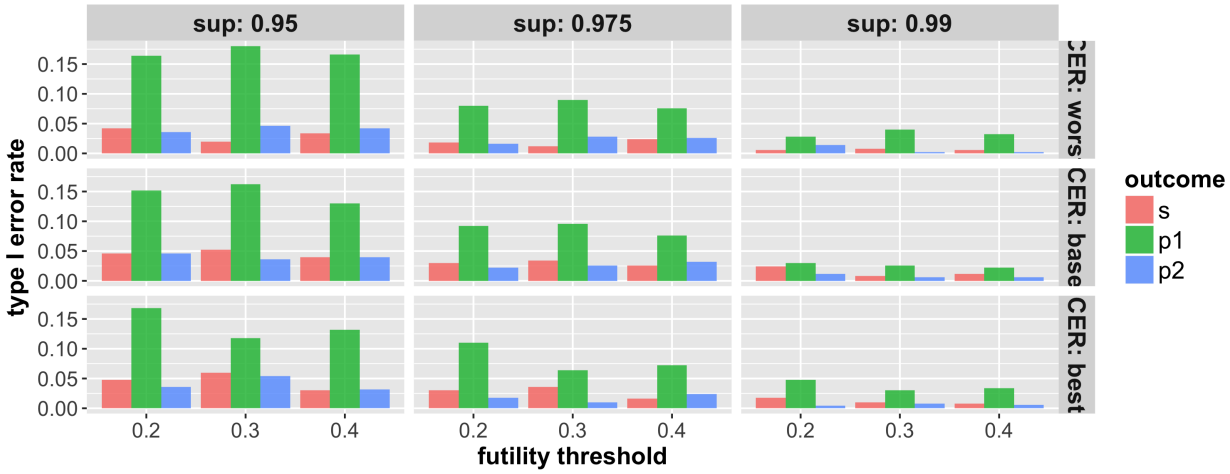
Stopping with moderately permissive outcome (p1)

Both superiority and futility decisions were made with respect to the moderately permissive outcome (p1).

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. 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. Specifically, the type I error for the three superiority thresholds 0.95, 0.975, 0.99 for these outcomes are approximately 5%, 2.5%, and <1%. For outcome p1, the type I error is approximately 15%, 10%, and 5% for the three superiority thresholds. As such, the 0.95 and 0.975 superiority thresholds do not appear to retain adequate control of the type I error for p1.

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 three superiority thresholds are presented by columns. Results by the futility thresholds are presented by the x-axis within each sub plot.



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, with only a moderate reduction in power when the true RRR=20%. 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=20%. For RRR=20%, the power to detect p1 generally decreases as the superiority threshold increases and the CER goes from best

to worst. 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 estimated for superiority thresholds are presented by the sub-plot x-axis. Power estimates by the futility thresholds are presented by the bar color (see legend).

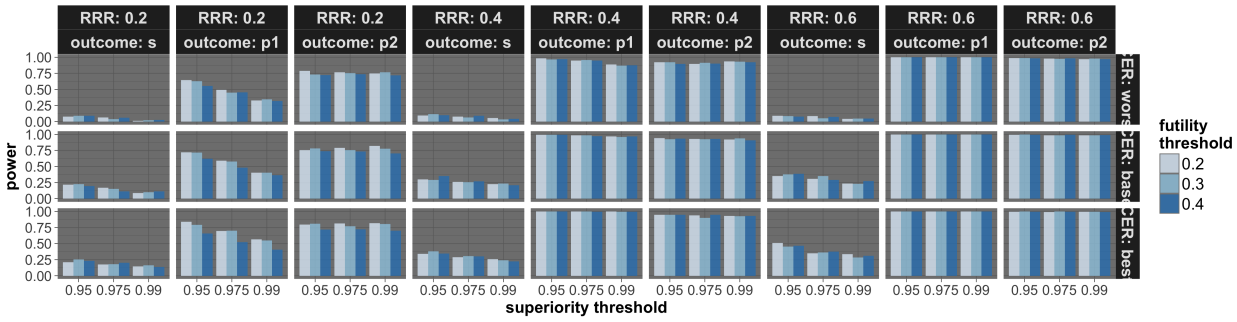
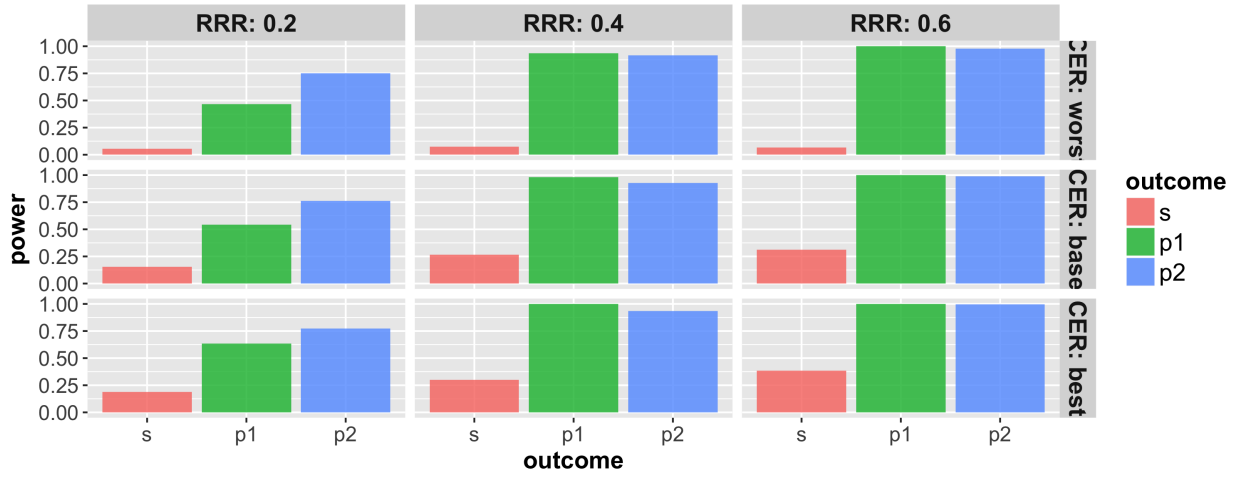


Figure 3. Under the true RRR=40% and 60%, power for outcomes p1 and p2 is very high but decreases to roughly 50-75% for the true RRR=20%. Power for outcome s is low for all scenarios, reaching roughly 25% for a true RRR=60% and the best case CER.

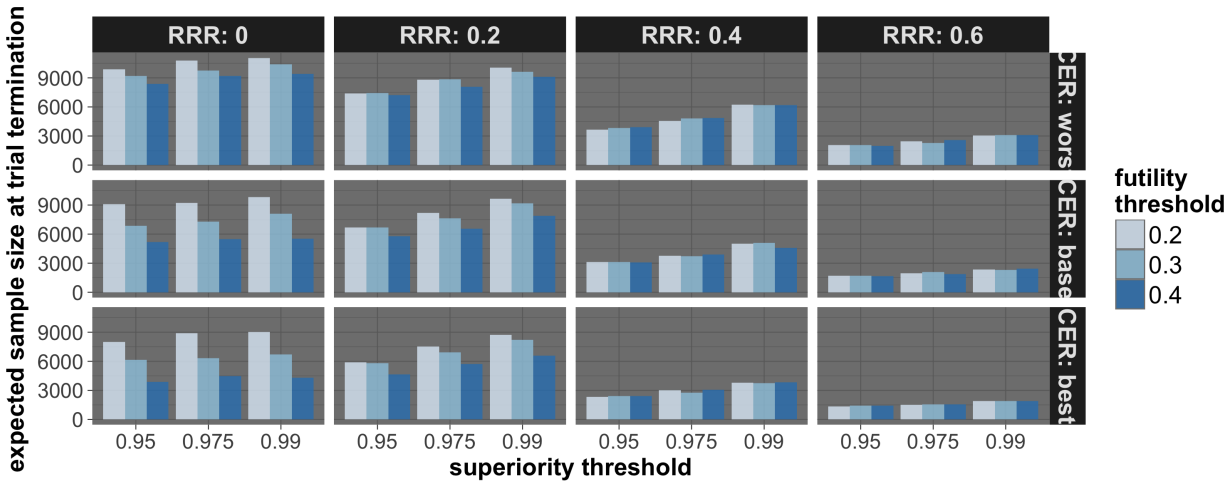
Figure 3: Power under the three relative risk reductions (RRR - upper column label) for each of the outcomes (lower column labels or legend). Observed power for each outcome is presented by colored bars.



Expected Sample Size

Figure 4 presents the expected (mean) sample size at trial termination. For true RRR=0 and RRR=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). Within these RRR, greater superiority thresholds only marginally increased expected sample size, but futility thresholds had negligible effects within each superiority threshold.

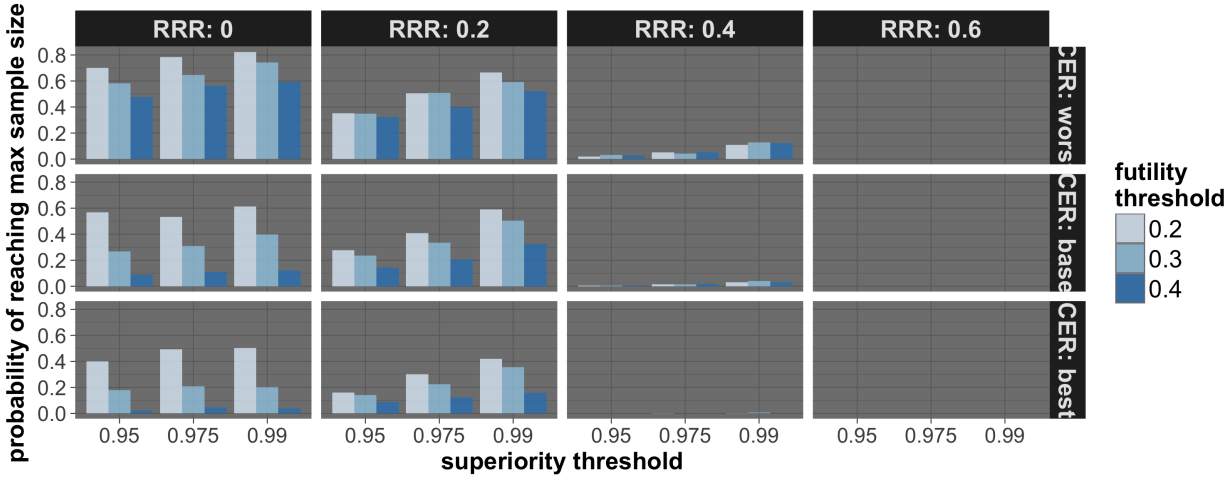
Figure 4: Expected sample size at trial termination. Results by control event scenarios are presented by rows. Results by relative risk reductions are presented by columns. Results by superiority thresholds are presented by the x-axis of the sub-plots, and results by futility thresholds are presented 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 and RRR=20%, and generally decreased as the CER improved. For RRR=40% and RRR=60%, this probability was negligible for all cases except RRR=40% and the worst case CER scenario, where it was roughly 10% for the highest superiority threshold.

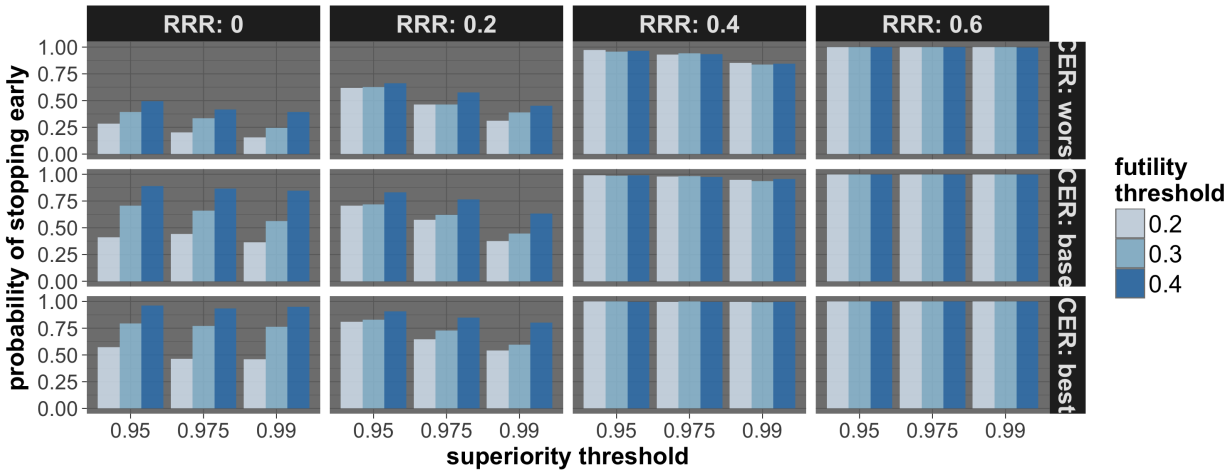
Figure 5: Probability of reaching maximum sample size for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), three superiority thresholds (x axis) and three futility thresholds (legend).



Probability of Stopping Early

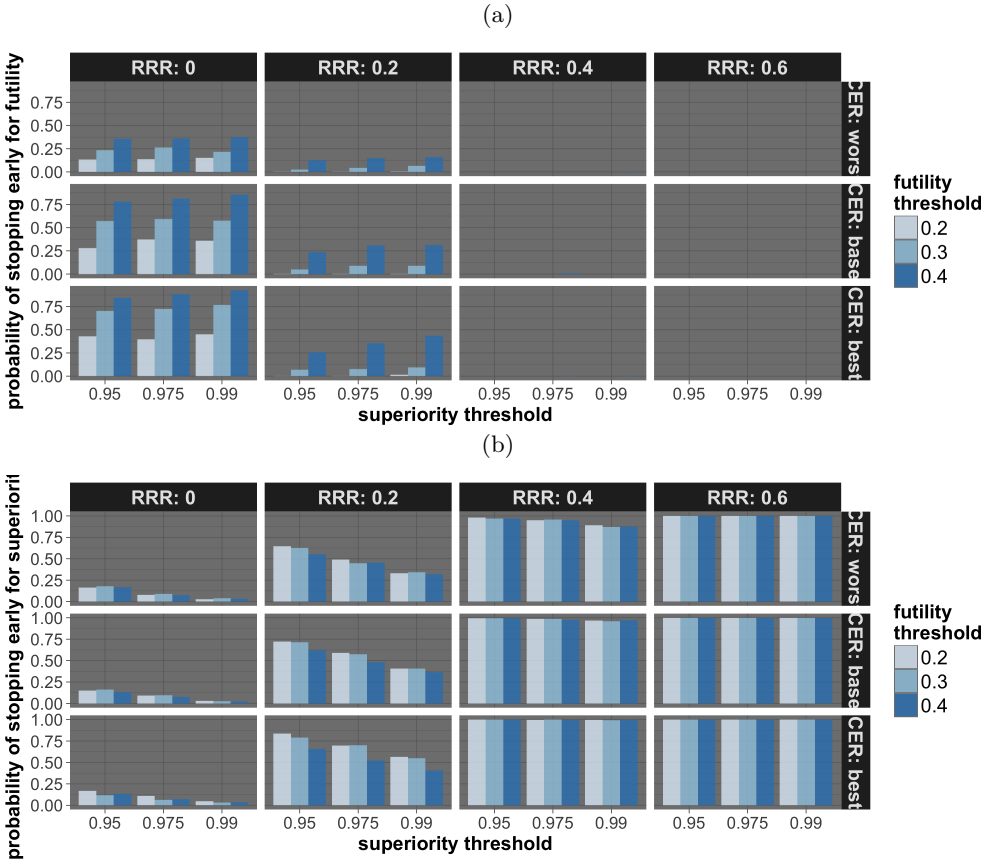
The probability of stopping early was obtained overall, for superiority and for futility. Figure 6 displays the overall probability of stopping early, Figure 7(a) displays the probability of stopping early for futility and figure 7(b) displays the probability of stopping early for superiority. The overall probability of stopping early is strongly positively correlated with the CER and RRR. When the simulated RRR=0% 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% and 40%, 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), three superiority thresholds (x axis) and three futility thresholds (legend).

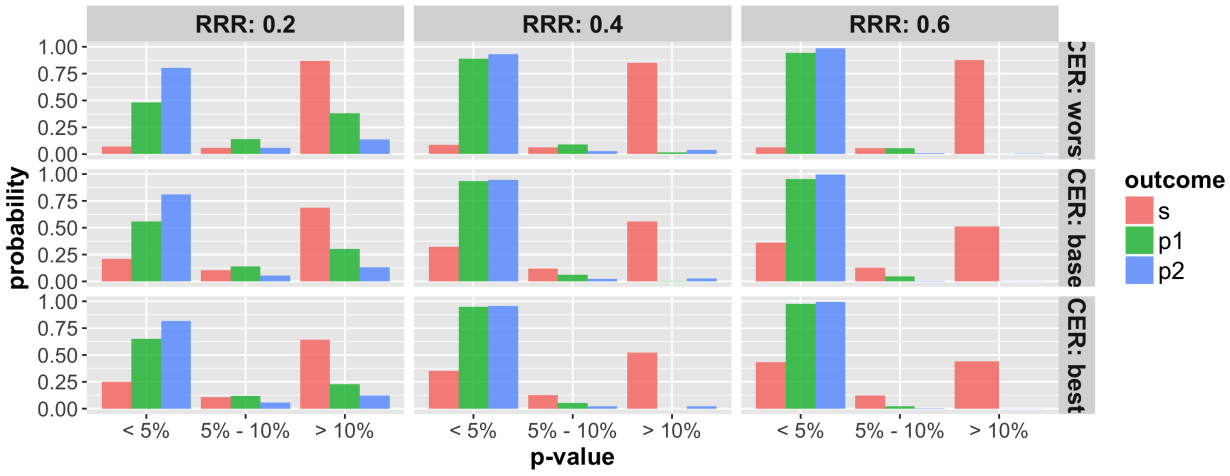


P-values at trial termination when a true effect exists

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. For the true RRR=20%, p-values less than 5% are found in roughly 75% of scenarios for p2, roughly 50% for p1 and 5-25% of s as the CER improves. A large proportion of these scenarios found p-values >5% for outcome s.

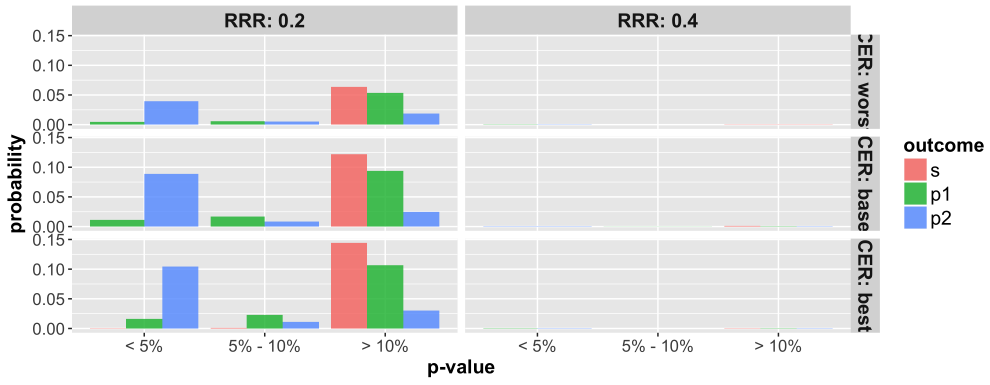
Figure 8: Overall probability at trial termination that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10%. The rows represent the three control even rate scenarios and the three columns present the three relative risk reduction scenarios.



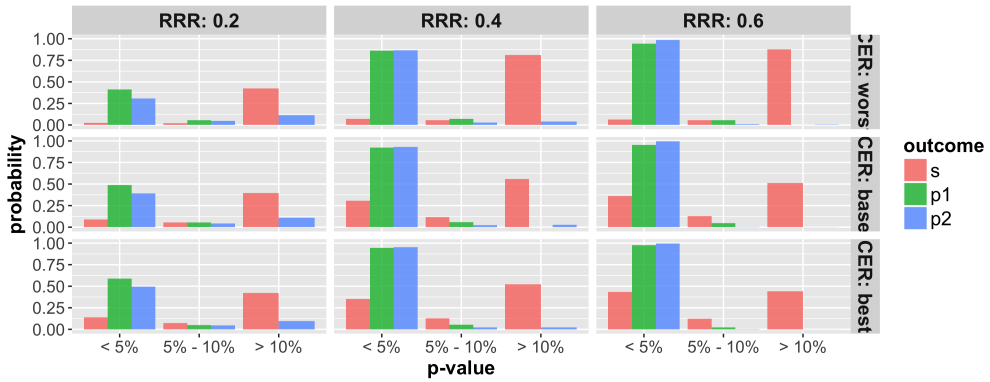
In the situations where the trial is stopped early for futility (shown here for only true RRR=20% and 40%), the majority of p-values for outcomes p2 are smaller than 5%, but are greater than 10% for outcomes p1 and p2 when the true RRR=20%. This and the small counts for RRR=40% can be explained by the 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)



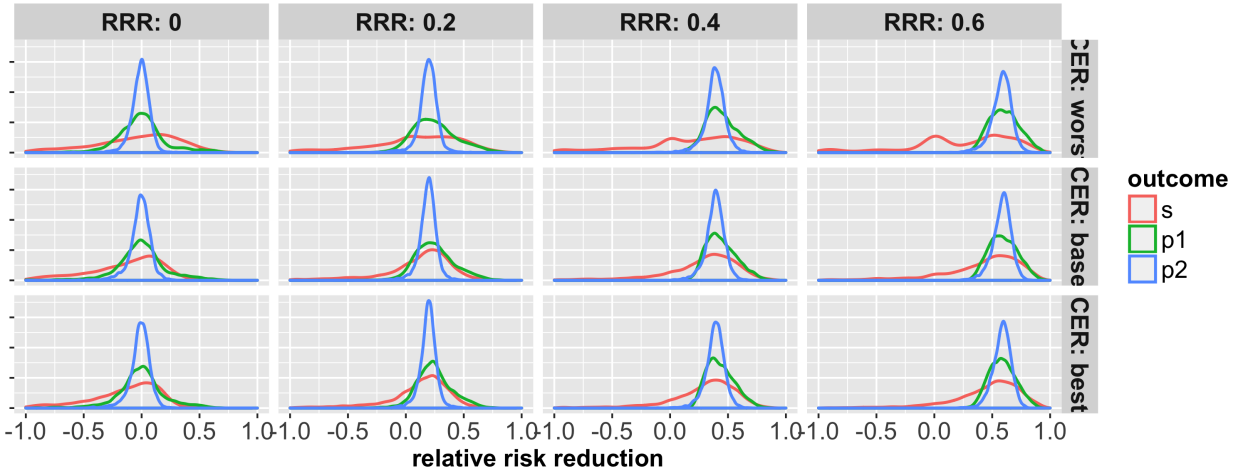
(b)



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.

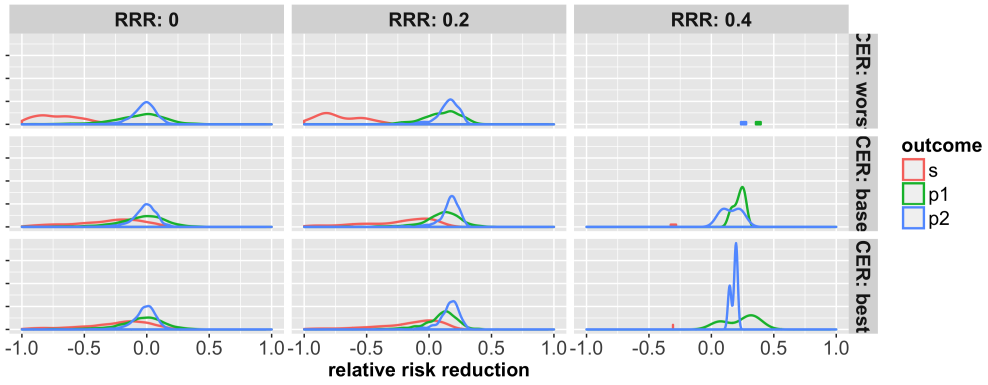
Figure 10: Distribution of relative risk reduction estimates (smoothed by a kernel density estimator) for the three control event rates (CER – rows), four relative risk reductions (RRR – columns) and the three outcomes (legend).



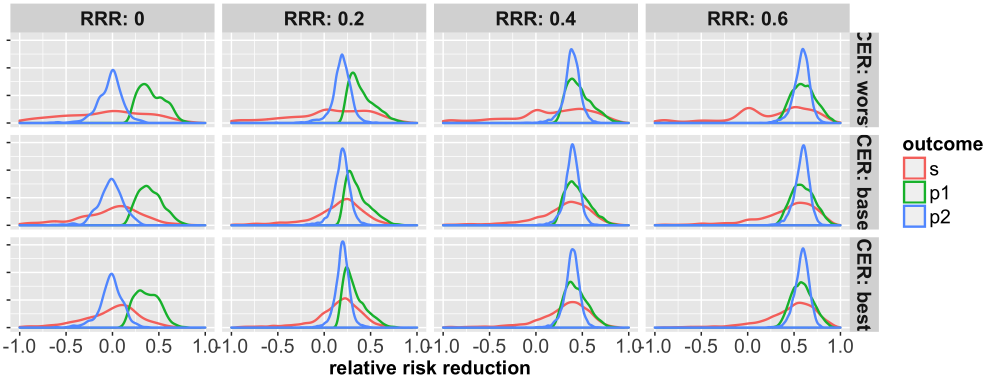
When stopping for futility, outcome s incurs small to moderate downward bias for all scenarios. For true RRR=0 and 20% and across all CER scenarios, both outcomes p1 and p2 show negligible bias, though p2 shows greater precision. For RRR=40%, outcome p1 shows less downward bias than p2 though both are imprecise with multiple peaks. 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 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)



Multi-arm trial: B.I., L.P., and Control

Note that pairwise comparison with control is employed and that the three arms are not compared simultaneously. The superiority threshold for L.P. is kept fixed at 0.99 and results are reported for the superiority thresholds of 0.995 and 0.999 for B.I. The futility bound was kept fixed at 0.2 and a probability threshold of 0.99 was required to stop for futility. Both superiority and futility decisions were made with respect to the moderately permissive outcome (p1).

Expected Sample Size

Figure 12 presents the expected (mean) sample size at trial termination. For true RRR=0 and RRR=20%, expected sample sizes were consistently high. For true of RRR=40% and RRR=60%, expected sample sizes were lower and decreased as the CER improved (worst to best). Within these RRR, the greater superiority threshold for B.I. only marginally increased expected sample size.

Figure 12: 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 superiority thresholds for B.I. are presented by the color of the bars (see legend).

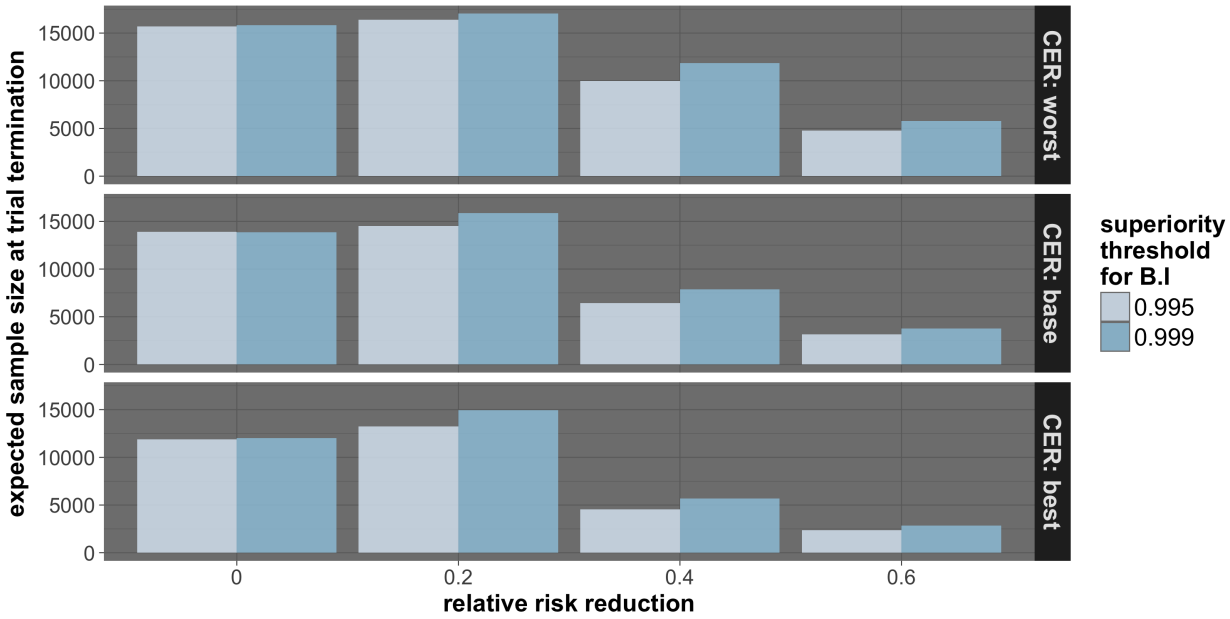
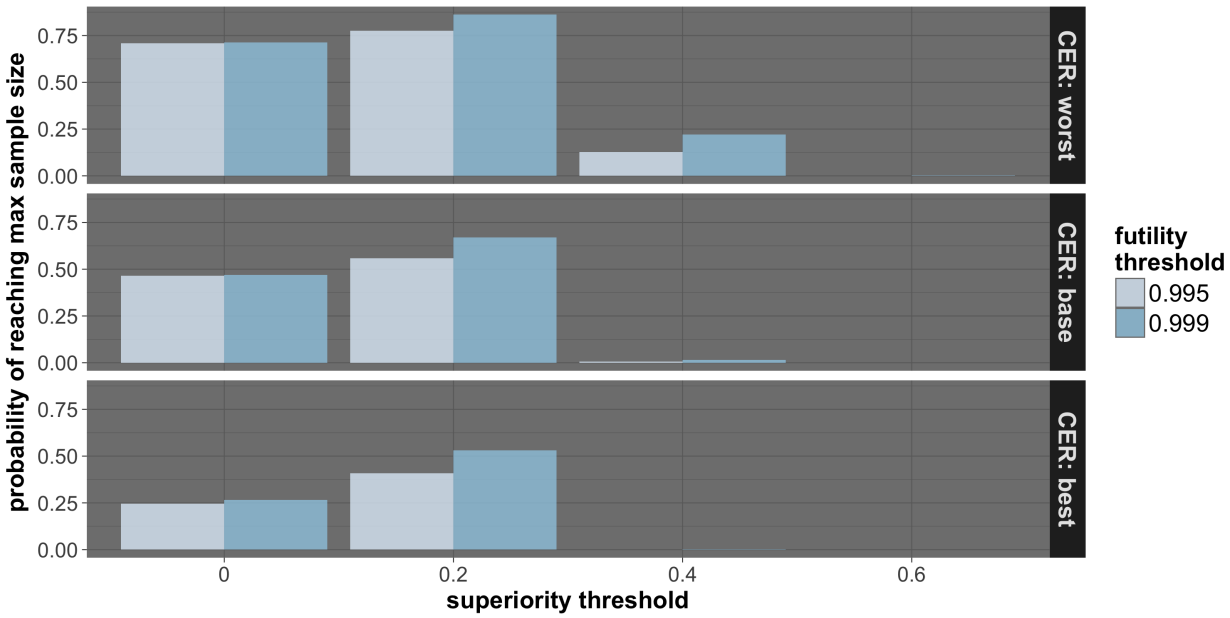


Figure 13 shows the probability of reaching the maximum allowed sample size of the trial. This probability was moderate to high for the true RRR=0 and RRR=20%, but decreased for the best case CER scenario. For RRR=40% and RRR=60%, this probability was negligible for all cases except RRR=40% and the worst case CER scenario, where it was roughly 25% for the highest superiority threshold.

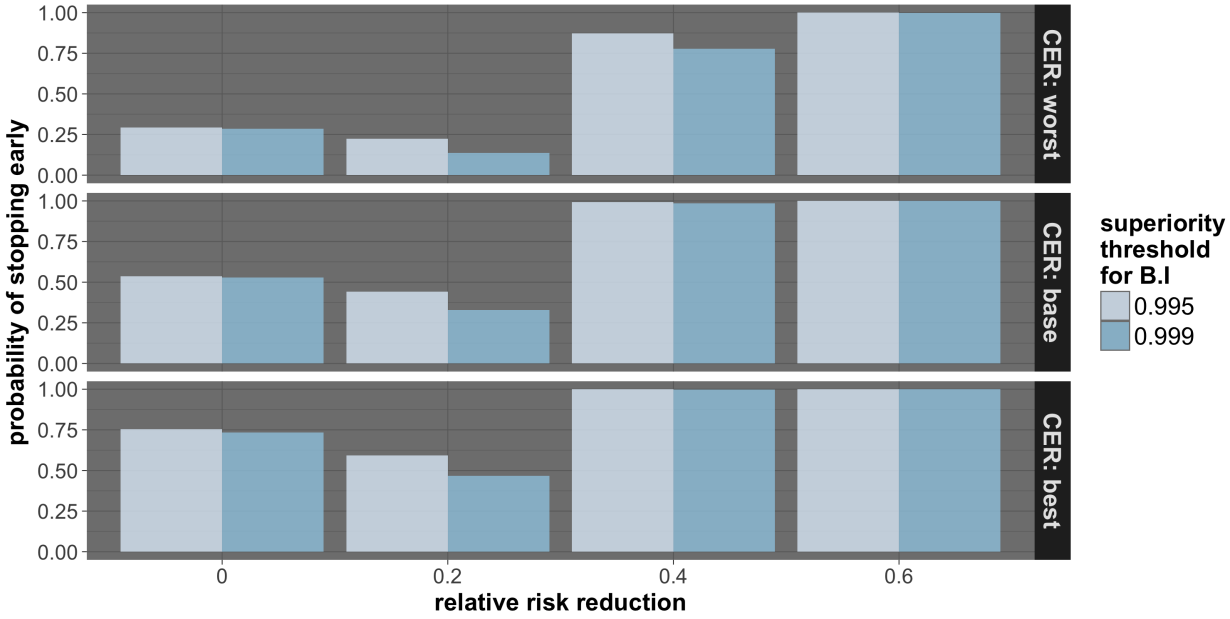
Figure 13: Probability of reaching maximum sample size for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and two superiority thresholds for B.I. (legend). Note that the legend and x axis labels should reflect the description given above, not as labeled on the graphic.



Probability of Stopping Early

The probability of stopping early was obtained overall, for superiority and for futility. Figure 14 displays the overall probability of stopping early, Figure 15 displays the probability of stopping early for futility ((a) B.I., (b) L.P.) and figure 16 displays the probability of stopping early for superiority ((a) B.I., (b) L.P.). The overall probability of stopping early is strongly positively correlated with the CER and RRR, being nearly certain for RRR=40% and RRR=60%.

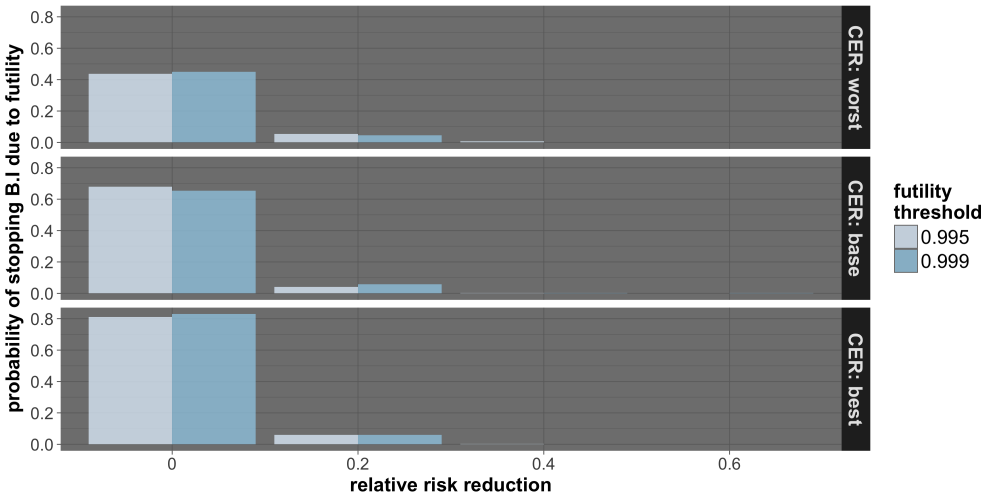
Figure 14: Probability of stopping early for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and two superiority thresholds for B.I. (legend).



Under the true RRR=0, there is a moderate to high chance of stopping for futility for B.I. or L.P. (the chance increases as the CER improves). There is negligible chance under all other scenarios. For B.I., there is a high chance of stopping for superiority for true RRR=40% and 60% and under all CER scenarios. There is negligible chance of stopping for superiority for RRR=0%, but a small to moderate chance (25-50%) under RRR=20%. Stopping for L.P. follows a similar trend for stopping for futility but shows a slight improvement over B.I. in stopping for superiority under the true RRR=20%.

Figure 15: Probability of stopping (a) B.I. or (b) L.P. due to futility. Stopping probabilities are presented for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and two superiority thresholds for B.I. (legend). Note that the legend and x axis labels should reflect the description given above, not as labeled on the graphic.

(a)



(b)

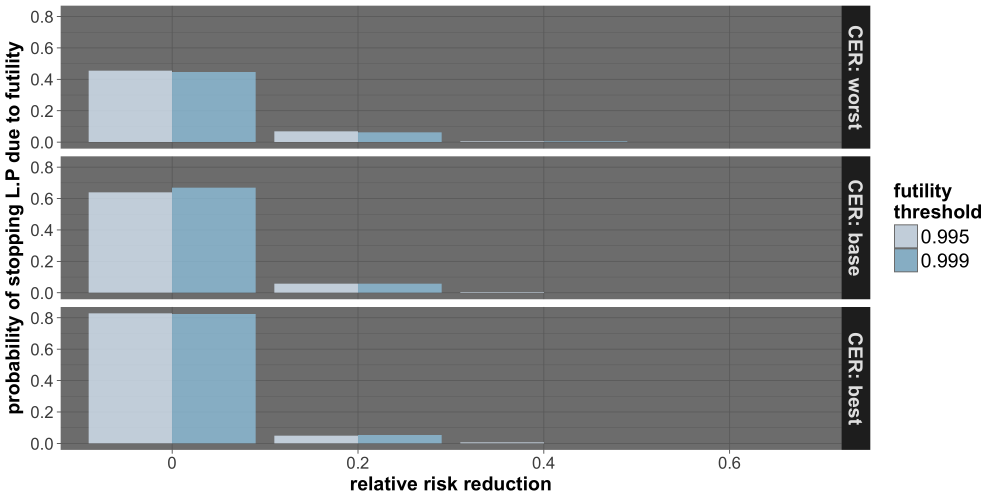
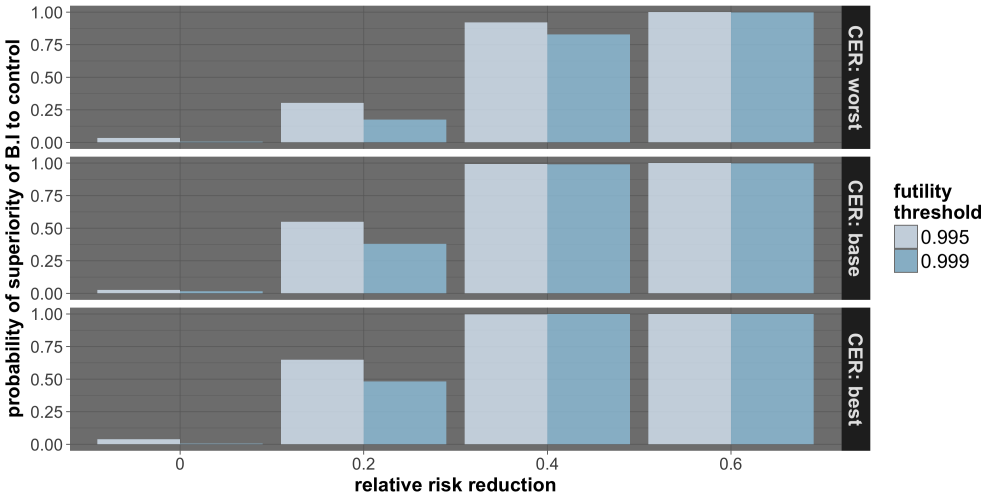
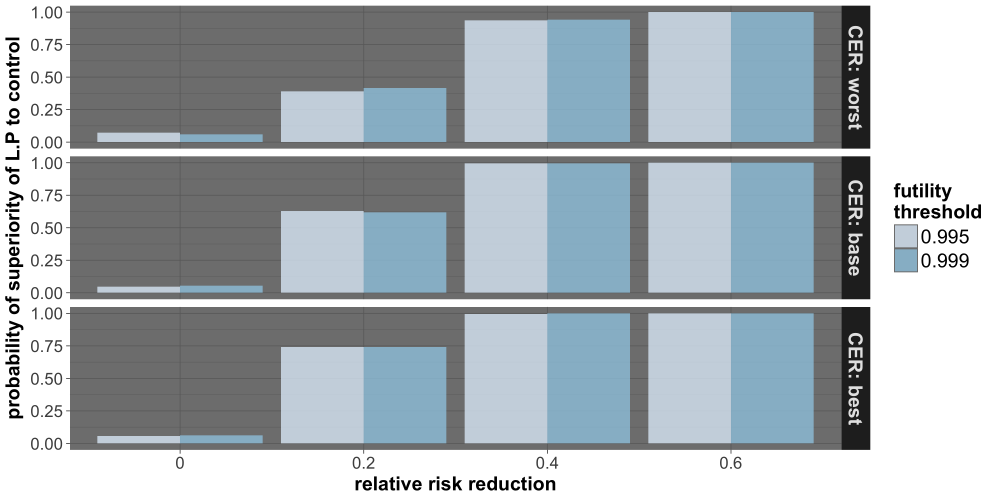


Figure 16: Probability of stopping (a) B.I. or (b) L.P. for superiority. Stopping probabilities are presented for the three control event rates (CER – rows), four relative risk reductions (RRR – columns), and two superiority thresholds for B.I. (legend). Note that the legend and x axis labels should reflect the description given above, not as labeled on the graphic.

(a)



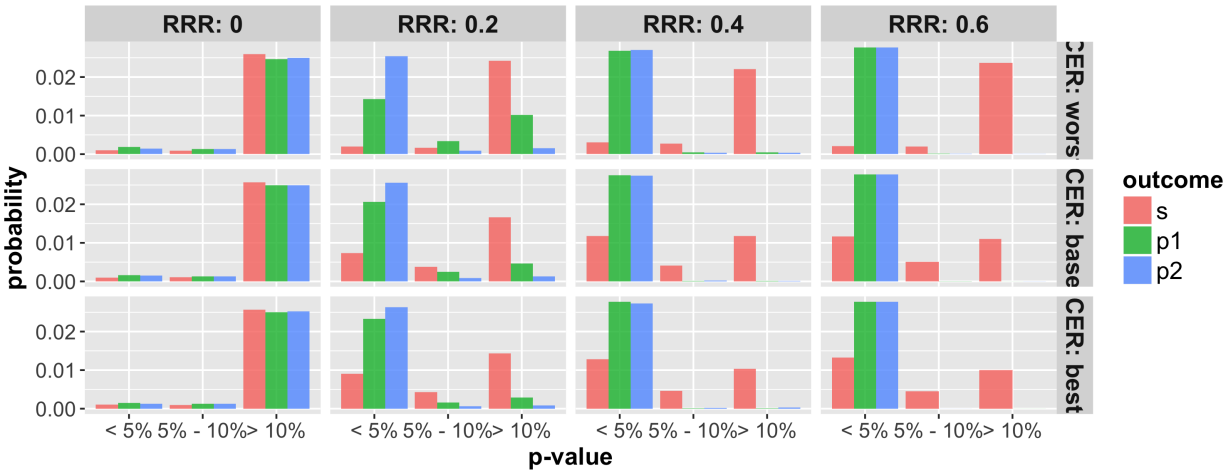
(b)



P-values at trial termination when a true effect exists

Figure 17 presents the categorical distribution of p-values (<5%, 5-10%, or >10%) upon trial termination (stopping early or reaching max. allowed sample size). Figure 18 presents the divide of p-values when stopping for futility ((a) B.I., (b) L.P.) and Figure 19 for and superiority ((a) B.I., (b) L.P.). 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. As the true RRR decreases, a higher proportion of p-values are >10% across all scenarios and outcomes.

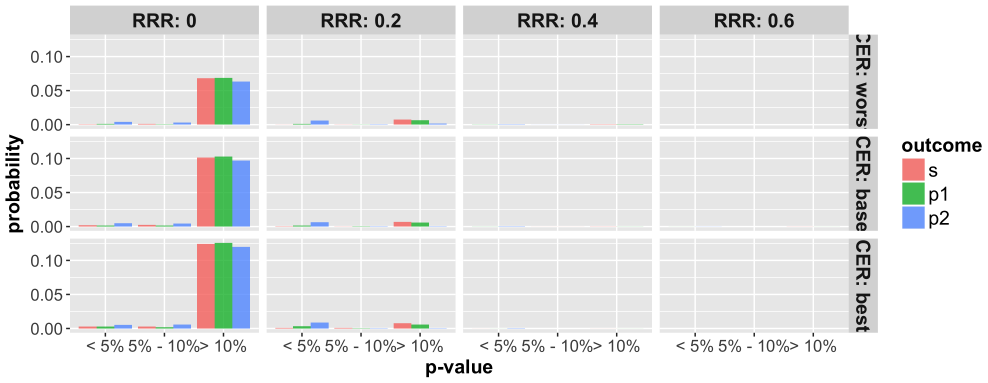
Figure 17: Overall probability at trial termination that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10%. The rows represent the three control even rate scenarios and the four columns present the four relative risk reduction scenarios.



In the situations where the trial is stopped early for futility (Figure 18) under the true RRR=0, the majority of p-values for all outcomes are greater than 10%. Reflecting that stopping for futility is less likely under the other scenarios, only a small number of p-values are found for these cases. When stopping for superiority (Figure 19) for both B.I. and L.P., a high proportion of p-values for outcomes p1 and p2 are less than 5% under true RRR=40% and 60%. A high proportion of p-values for outcome s are greater than 10% for all scenarios.

Figure 18: 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 stopped for futility of (a) B.I or (b) L.P. The rows represent the three control even rate scenarios and the four columns present four relative risk reduction scenarios. Note: In the figures below, the denominator in each figure is the number of simulations (not the number of trials stopped for futility or superiority, and thus, the proportions do not add up to 100% within one figure. Further, the figures do not include simulations where the trial went to the maximum allowed sample size. The bars should be interpreted with respect to the relative proportion that fit in each category.

(a)



(b)

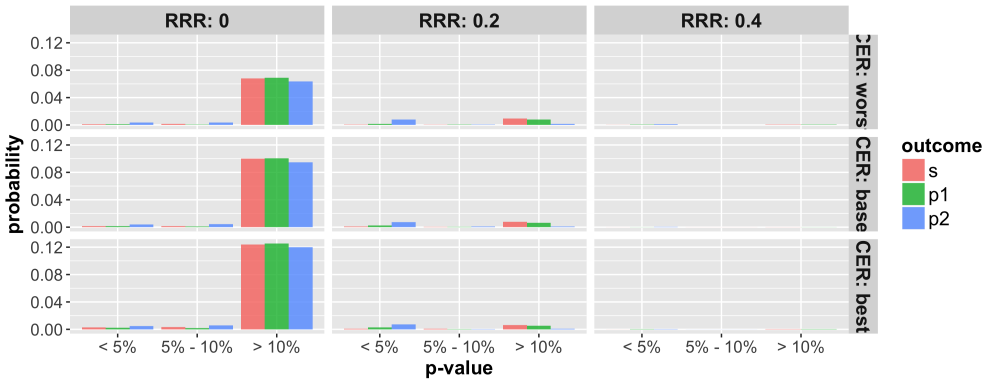
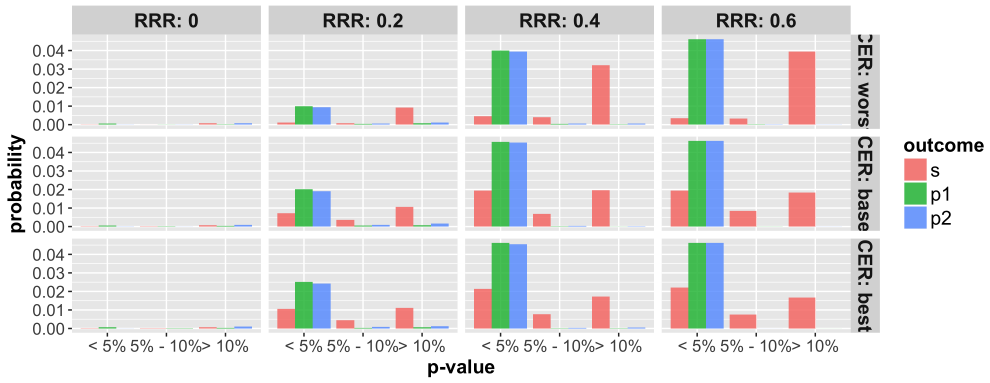
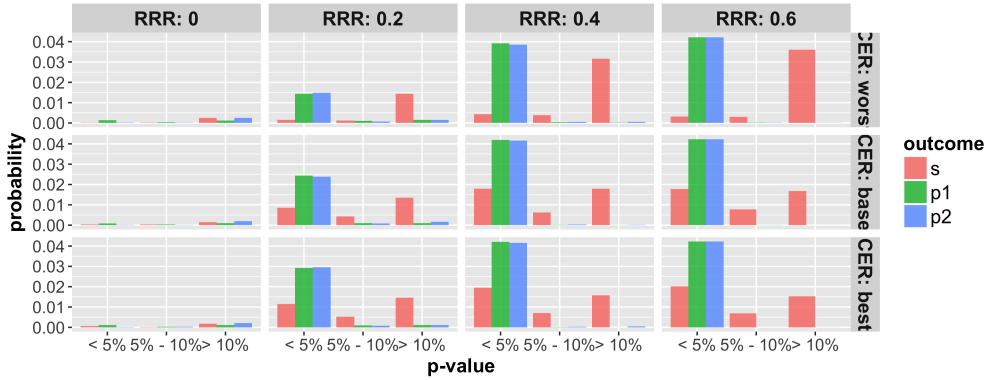


Figure 19: Probability that the p-value (from Fisher's exact test) at termination of the trial is below 5%, between 5% and 10% and greater than 10% for cases where trial was stopped for superiority of (a) B.I or (b) L.P. The rows represent the three control even rate scenarios and the four columns present four relative risk reduction scenarios. Note: In the figures below, the denominator in each figure is the number of simulations (not the number of trials stopped for futility or superiority, and thus, the proportions do not add up to 100% within one figure. Further, the figures do not include simulations where the trial went to the maximum allowed sample size. The bars should be interpreted with respect to the relative proportion that fit in each category.

(a)



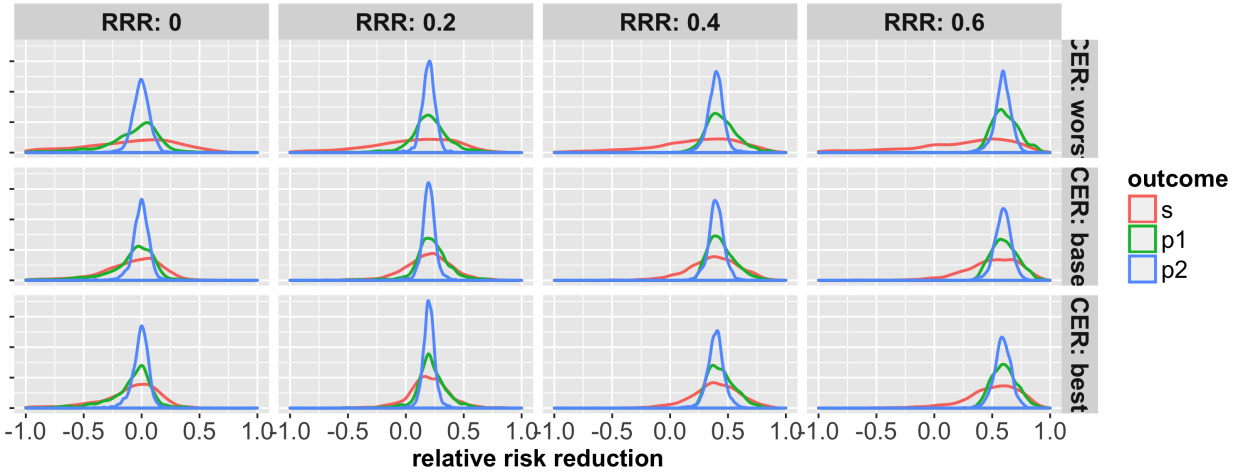
(b)



Relative risk reduction estimates at trial termination

Figure 20 presents the distribution of relative risk reduction estimates upon trial termination. Figure 21 ((a) B.I., (b) L.P.) and Figure 22 ((a) B.I., (b) L.P.) present the distribution of relative risk reduction estimates from trials stopped early for futility and superiority, respectively. In general, the estimates of p1 and p2 exhibit much larger precision.

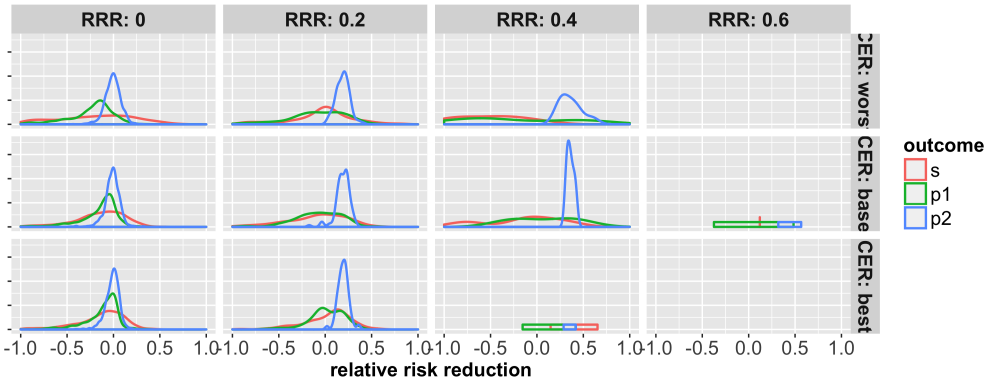
Figure 20: Distribution of relative risk reduction estimates (smoothed by a kernel density estimator) for the three control event rates (CER – rows), four relative risk reductions (RRR – columns) and the three outcomes (legend).



When stopping for futility (Figure 21), outcome p2 is generally precise and shows little bias. Outcomes p1 and s exhibit either small downward bias for lower RRR or become overly diffuse for higher RRR. When stopping for superiority, outcome p2 show little evidence of bias and has good precision across all scenarios. Outcomes p1 and s have small upward bias for lower RRR and show improvement in precision for higher RRR.

Figure 21: Distribution of relative risk reduction estimates after stopping early for futility of (a) B.I. or (b) L.P. Results are presented for the three control event rates by rows, four relative risk reductions (by columns) and the three outcomes (legend).

(a)



(b)

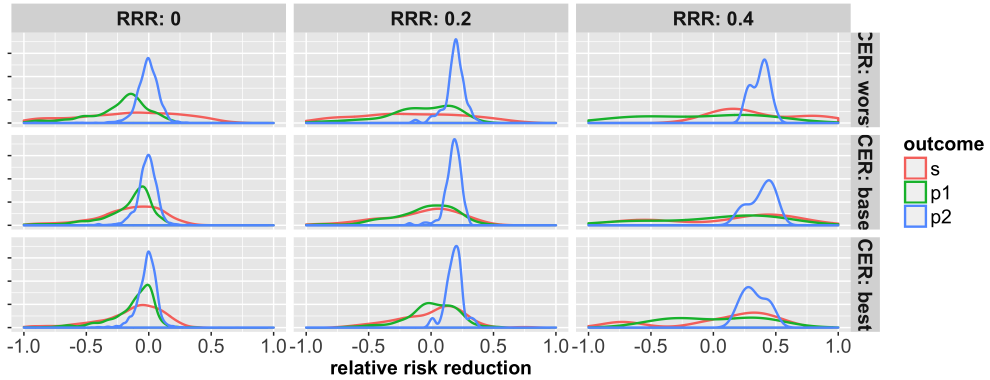
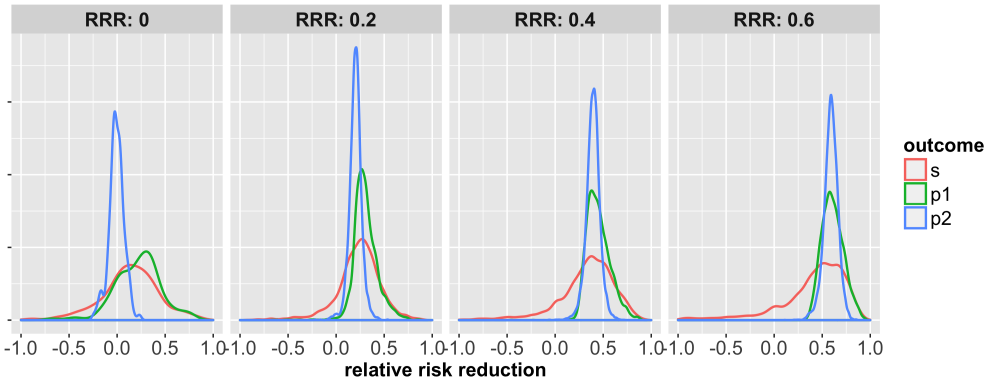
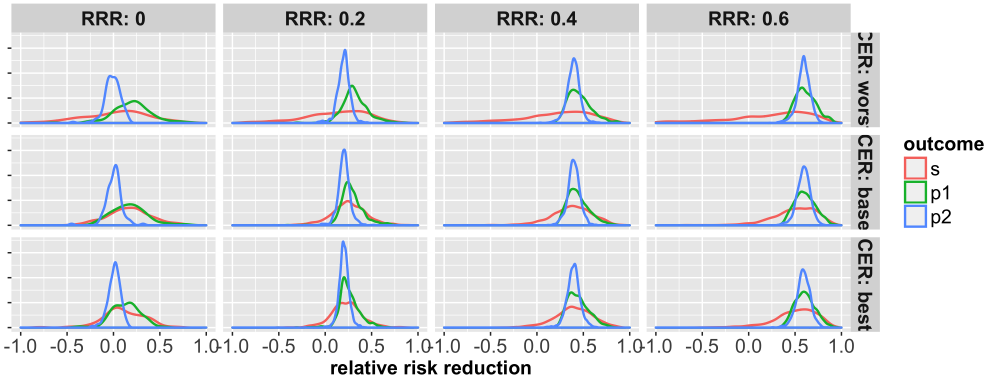


Figure 22: Distribution of relative risk reduction estimates after stopping early for superiority of (a) B.I. or (b) L.P. Results are presented for the control event rate (rows, BEST CASE ONLY for B.I?), four relative risk reductions (by columns) and the three outcomes (legend).

(a)



(b)



Summary

For the two arm trial with stopping rules for the moderately permissive outcome (p1), a superiority threshold of 0.99 should be employed to ensure good control of Type 1 error (false positive rate less than 5%). Under this design, there is good power to detect outcomes p1 and p2 across most scenarios and a high proportion of p-values expected to be less than 5%. Additionally, there is little bias and good precision in the RRR estimates of outcomes p1 and p2.

The multi-arm trial comparing B.I., L.P. and Control had higher expected sample size at trial termination and a high proportion of significant p-values is expected for RRR=20%, 40%, and 60% for outcomes p1 and p2. There is good precision in the RRR estimates of outcomes p1 and p2 when stopping for superiority, though this degrades when stopping for futility.