TACTIC-R Simulation Work

Simon Bond

22/09/2020

Abstract

Introduction

TACTIC-Ris a trial of repurposed drugs in the severe COVID population, with Standard of Care as control and initial 2 other arms (Baricitinib, Ravulizumab) for the treatments. TACTIC-E is a trial of experimental drugs in an otherwise similar trial design. The endpoint is time to failure (a composite endpoint) up to 14 days of follow-up. The intention is to have interim analyses that allow treatment arms to be dropped and/or new treatments to be added. The rules for such adaptations were left rather flexible at the time of the protocol being developed in very short time. This report is an attempt to consider the risks and advise on how such decisions could be made.

A simulation approach was taken to estimate the long term operating characteristics of proposed decision rules under a variety of scenarios representing the true data generatng process and thus the treatment effects. This document has a double purpose: document how this was done with a skeleton guide to the code files that were developed ; present the results of the simulations and make recomendations. The R code files are mirorred in a github repository `shug0131\tactic`.

Outline of Simulations

The approach take breaks down into 3 broad steps within any particular scenario

- Generate data up to the first interim, and produce the interim analyses. This was already summarised in a previous report, but will be recapitulated in this document
- generate future data over a range of sample sizes, and apply off-the-shelf stopping rules
- repeat the preceding steps 1000s of times, and then summarise the key outputs to estimate the opearting characteristics (biases, type 1/2 errors, coverage, mean squared error)

The computing resources needed to carry this out in a timely fashion were substantial. Through the MRC Biostatistics Unit, discretionary access was provided to the High Performance Computing (HPC) facility at Cambridge University, which was utilised to perfrom Array parallel processing for the repetitions described above.

Interim Analysis

Data Generation

The scenarios considered are described in the table below. Further scenarios coudl be added if desired.

Label	14-day Control Failure rate	Hazar ratio Bar v SoC	Hazard Ratio Rav v SoC	Drop out rate
one_winner	0.7	0.7	1	0.1
null	0.7	1	1	0.1

Label	14-day Control Failure rate	Hazar ratio Bar v SoC	Hazard Ratio Rav v SoC	Drop out rate
medium_events	0.8	0.7	1	0.1
low_events	0.9	0.7	1	0.1
worst_definition	0.8	0.7	0.85	0.1
good_ugly	0.9	0.6	0.85	0.1

The first two lines, correspond to the alternative and null hypotheses, with just one effective arm for the alternative. Medium_events and low_events, consider varying the assumption of the event rate, but not the HR. The last two have an arm that has a HR in a positive definition, but below what woudl be regarded as clinically significant; this is the most challenging scenario to discern which arm to continue or drop.

Data is generated according to the input parameters by functions defined in the script `sim_failure.R`, which incoporates censoring, dropouts and a random site effect with a SD on the log hazard scale of 0.1, over 10 sites, assuming an exponential distribution on the individual patient level.

Prior Distributions

For all the Bayesian analyses, three prior distributions were used, labelled as "Vague", "Pessimistic" and "Optimistic". They were all normal distributions on the log HR scale, set to match the 5%, 50% or 95% quantiles to HR as desired, anchored around the alternative hypothesis of HR=0.7, and the null HR=1, in the cases of the optimistic and pessimistic priors.

Prior	5%	50%	95%
Vague	0.5	-	2
Pessimistic	0.7	1	-
Optimistic	-	0.7	1

All other nuisance parameters were set to have a vague prior as provided by default by the software.

Analysis

The rstanarm R package was used to perform MCMC inference using the generated data and the priors. The Poisson trick was used to fit a mixed effects GLM poisson model with fixed effects for each time interval (days 1 to 14), treatment, and random site-level intercept. Given the small number of days, compared to the sample size, smoothing tecniques were not used to estimate the baseline hazard. The fitted model was saved under a file name of `fits_image_*.rds` where the `*` is an index number over the repeated simulations performed.

2 Chains of posterior sampels were provided, of length 1500 using thining, whereby alternatve observations were discarded to avoid auto-correlation.

Following on, the fitted models were used to calculate the posterior probability that each of the two HRs lay within several reference values (0.8,1,1.25), and the outputs saved as `posterior_probs_*.rds`.

A standard frequentist version of the fitted model from `coxph()` was also saved into `coxfit_*.rds`

Then we generate 1500 pseudo observations from the predictive posterior distribution, taking a bootstrap sample of baseline covariates as needed. This was iterated 1500 times (the same number as the posterior sampling), and then processed into a set of 1500 combined pseudo data sets of the original data and predictive posterior distribution samples `new_data.R`.

The sample size of each data thus exceeds 2000 patients. For each combination of arms to continue recruiting (all 3, dropping either of Bar or Rav), and a fixed range of sample sizes (458,687,938,1407), for the *overall sample size* regardless of there being 2 or 3 arms, a subset of the large data set was taken to match up. A standard frequentist cox model was fitted to each subset and each of the 1500 replications, from which the

estimates and SE of the HRs were recorded, along with a declaration of statisticial significance for either or both comparisons. These were then averaged across the replications to provide the predictive power. A data set with rows for each arm combination and sample size, plus the predictive power was saved as `power_*.rds` . However, currently the predictive power was not used to make decisions, in the subsquent simulations below. Future work could consider using this interim calculation to make choices on sample size or which arms to use.

Computation details.

This step was implemented by running the shell script `slurm_wrapper.sh`, which takes two input values, a string giving the name of the scenario to use, and the number of parallel cores to use in the R script. The script creates directories to save results named after the scenario, takes a copy of the R script, and runs `sbatch slurm_model_sim_fit_skylake.txt` . The template code `slurm_model_sim_fit_skylake.txt` for batch Array processing on the HPC cpu cluster impliments several steps:

- sets the number of replications to be 1000, creating 1000 array jobs.
- loads r version 3.6
- sets the run time to be a max of 36 hours
- runs `Rscript model_sim_and_fit.R $scenario $parallel` for each array. The last parameters are passed from the original input to `slurm_wrapper.sh`

Future Data

As distinct from generating data from the predictive posterior, which is the only option during the real trial at the interim stage, for the *simulation* we next generate future data according the true parameters as set by the choice of scenario.

Sufficient data is generated to cover all the possible combinations of arms and maximum sample sizes.

Stopping boundaries for Off-the-shelf group sequential designs, for the relevant interim sample sizes, were calculated within `group_sequential_design.R`, for O'Brien-Femming, Triangular, and Pocock error-spending. At each interim size a frequentist `coxph()` fit was performed and a bayesian model fitted as well. The stopping rules were evaluated, estimates recorded, and bayesian psoterior probabilities calculated much like at the initial interim, and a rule to stop an arm for success if $\Pr(HR<0)>0.95$, considered for each of the three prior distributions.

But no further calculation of the predictive power was not repeated, as the aim of this report and simulation exercise is to help judge what the optimal course of action is *at the first interim*. Conceptually, backward induction could be used to improve this, but at the moment would be too challenging computationally.

This was repeated to match up with the 1000 iterations of data orginally generated, and the outputs of the first interim analysis. For each iteration and stopping rule, a final analysis was performed to provide the estimate/SE, incidence of the 95% CI containing the true parameter, sample size at stopping; this was done using frequentist coxph() methods, and a bayesian version thereof, but only using the vague prior and excludign the other two priors for parsimony.

Each iteration save the set of results at `ifnerence_results_*.rds`

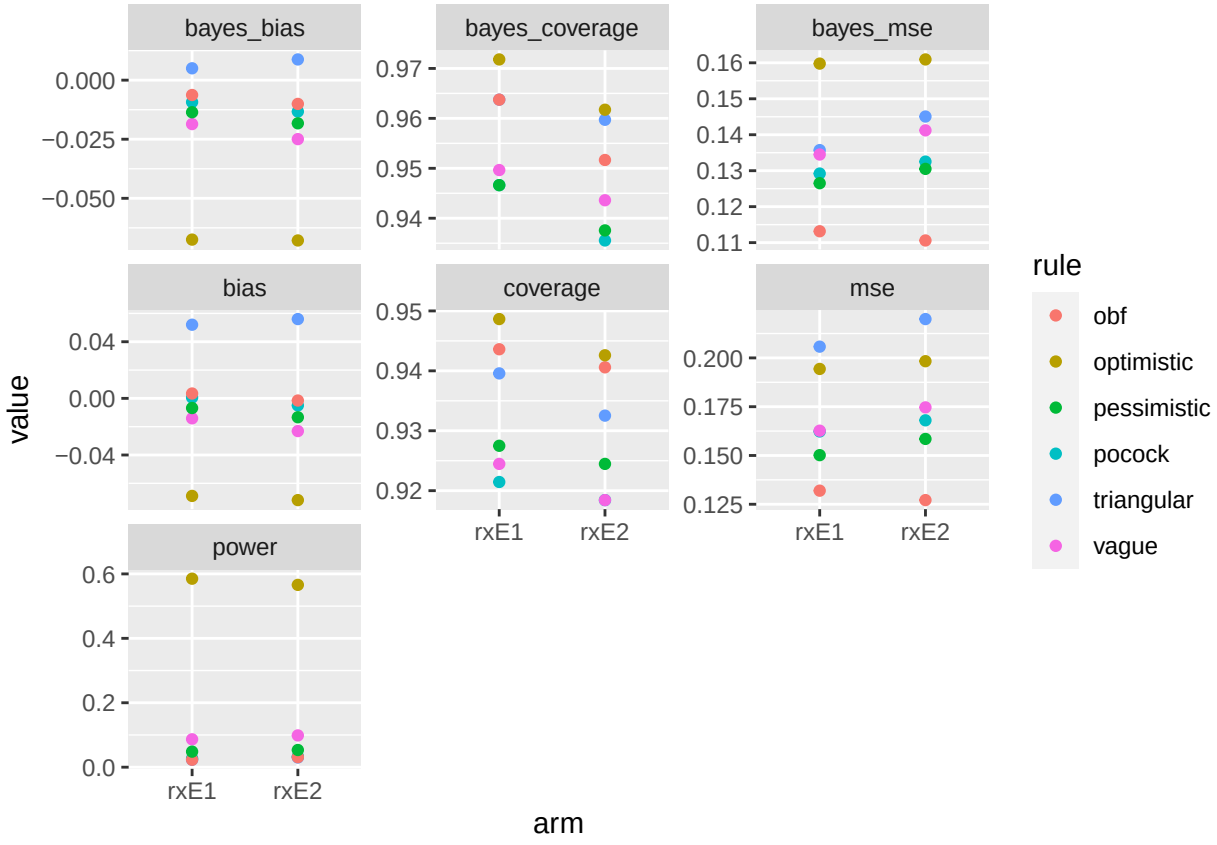
The final aggregation of results of the iterations calculates: bias, power, mean squared error, and coverage, applied to the frequentist and bayesian (excluding power) estimation within `operating_characteristics.R` with the results for each scenario in `oc_results.rds`.

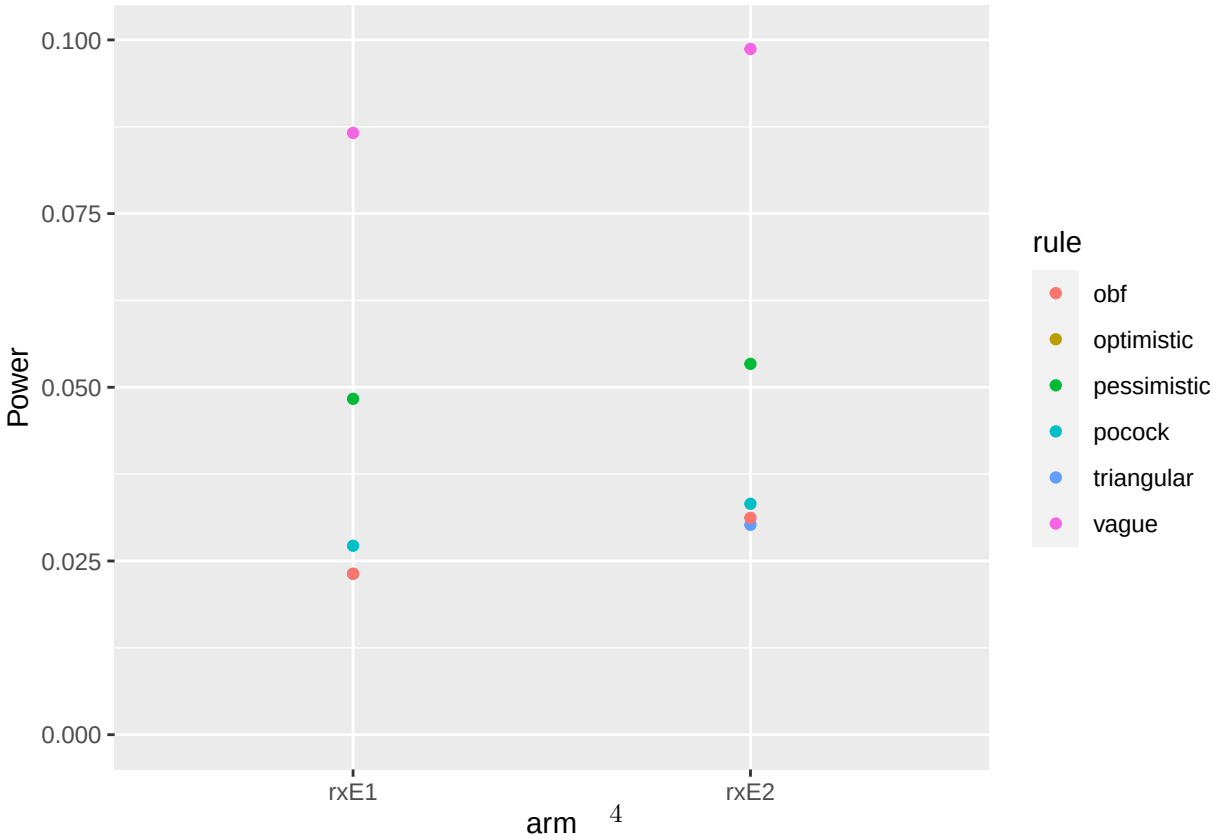
Computation Details

Similar to the first set of batch computing, we run the shell script `slurm_wrapper_future_inference "null" 1`, which calls the template `slurm_future_inference_skylake.txt`, whcih repeated calls `future_real_data.R` .

Results

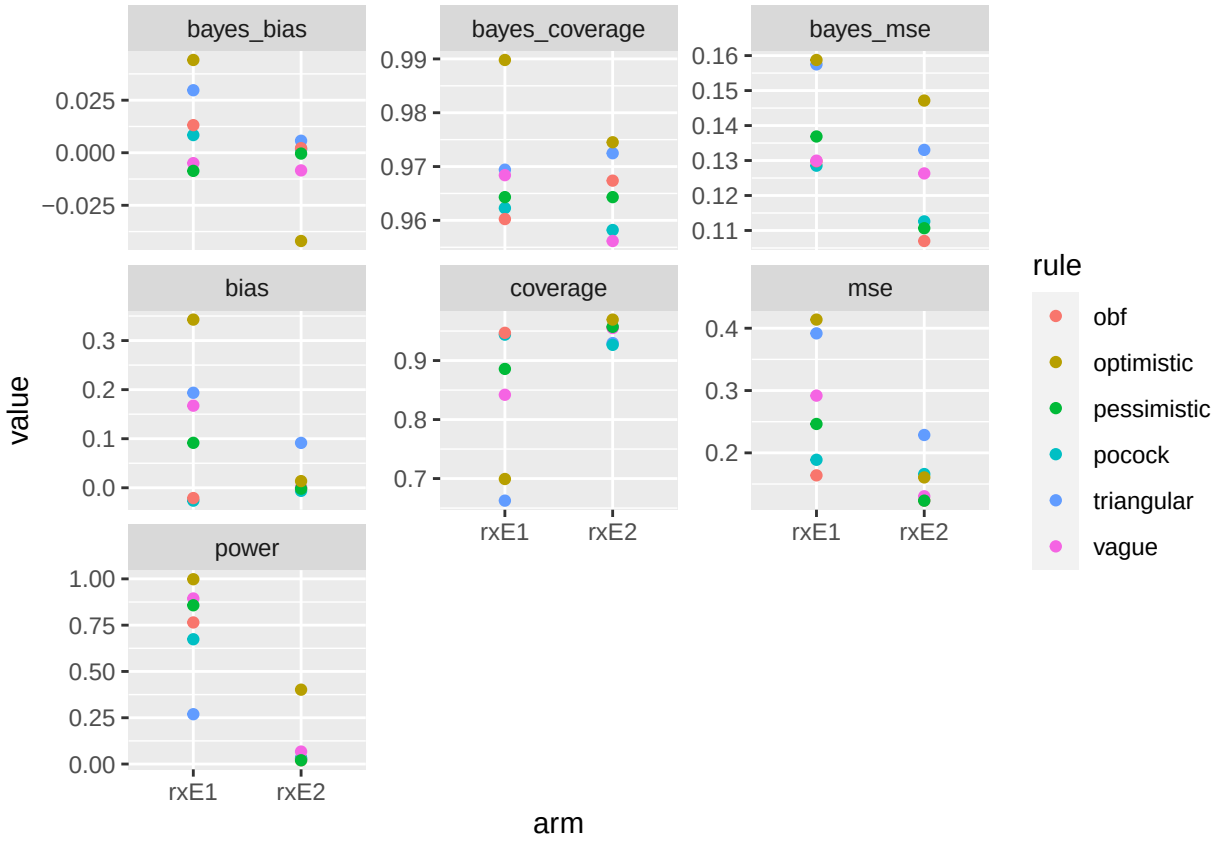
Null



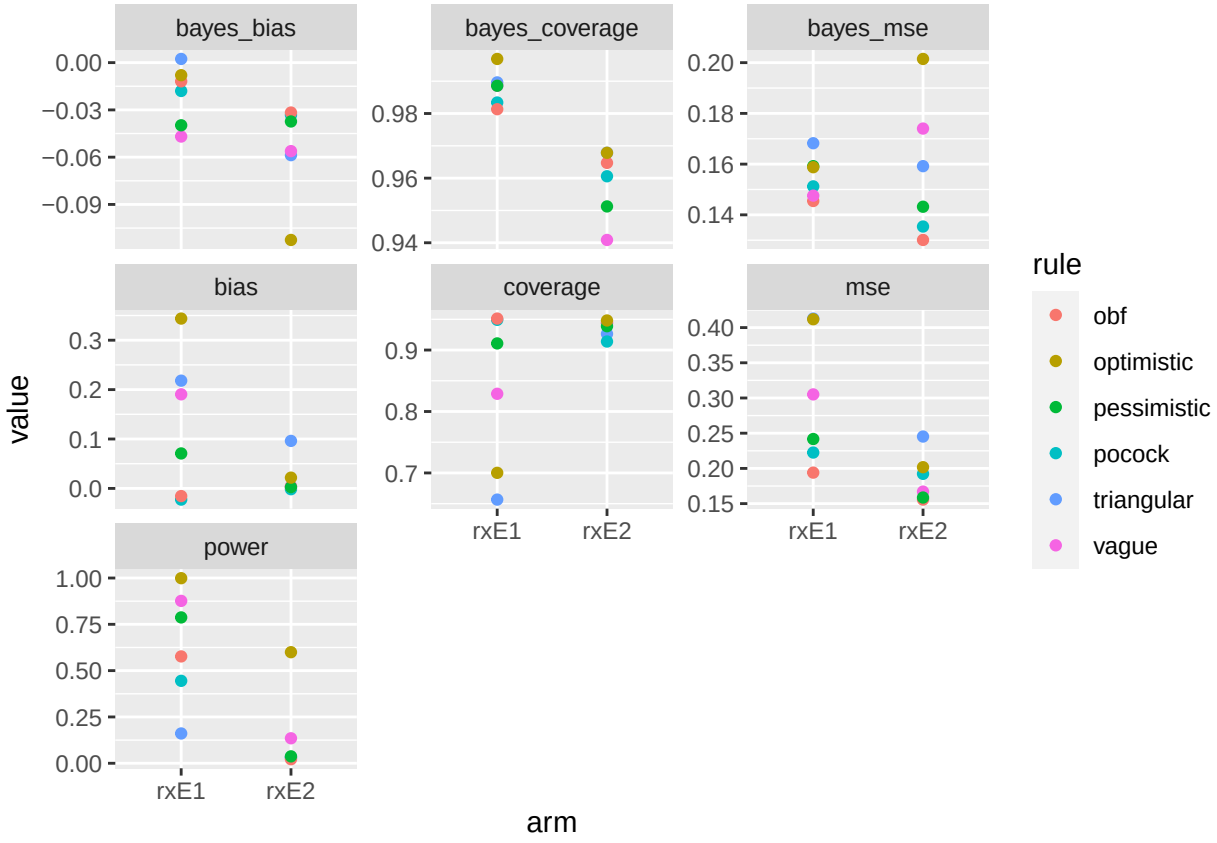


Alternative HR=0.7 & 1

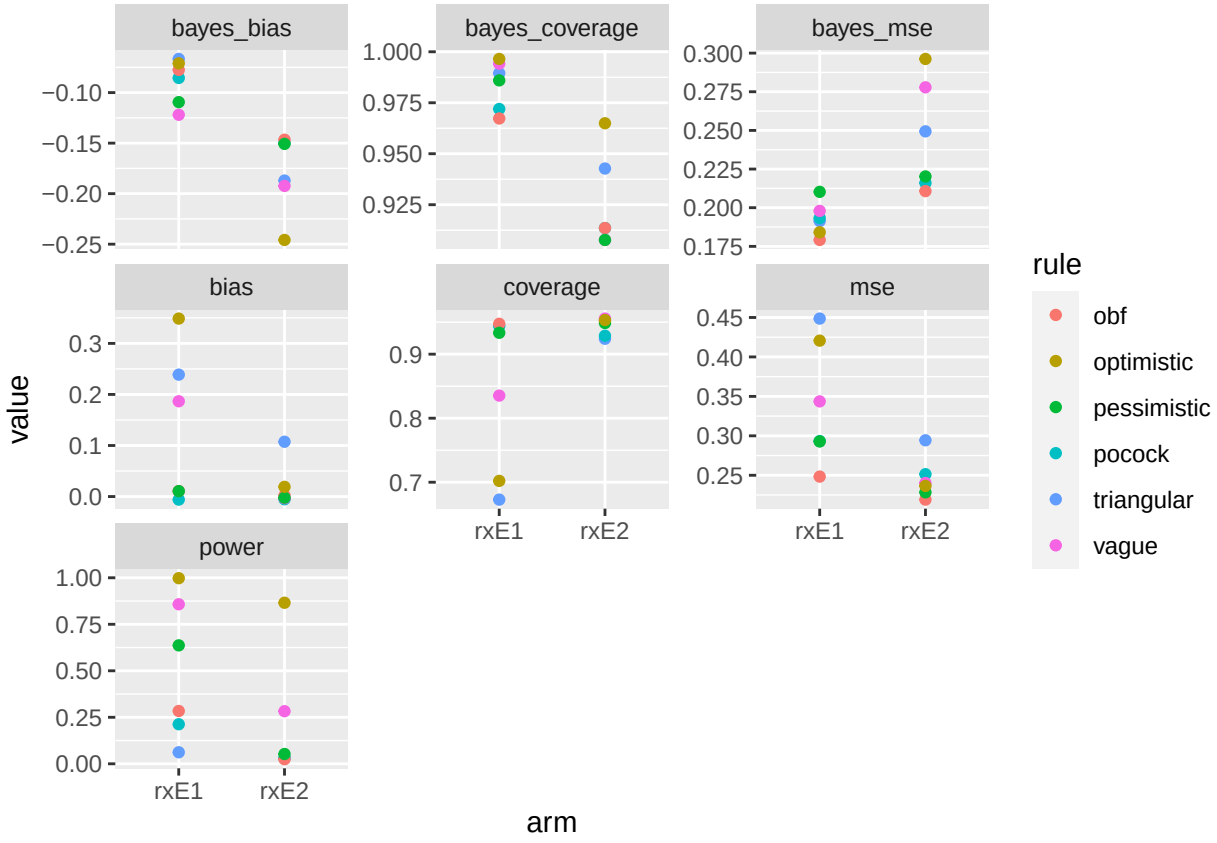
Control Rate = 0.7



Control Rate = 0.8

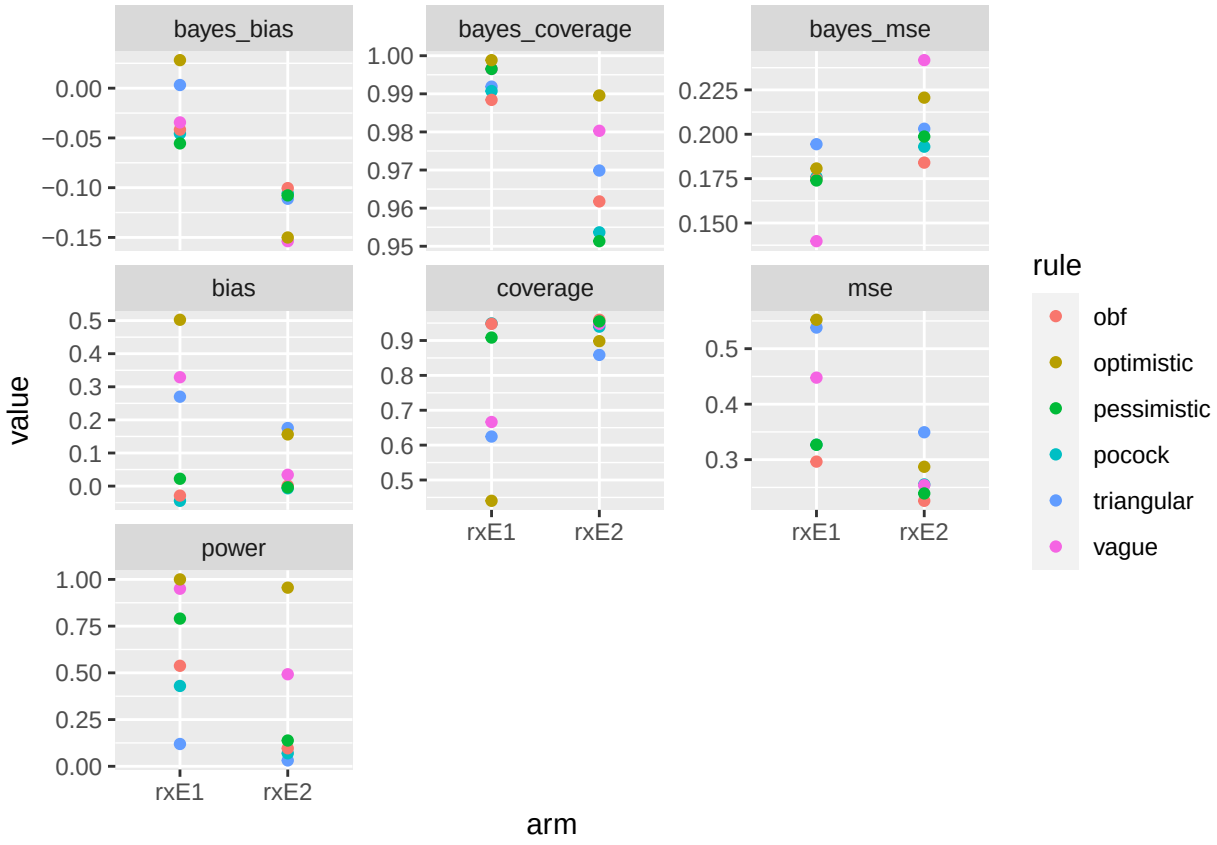


Control Rate = 0.9

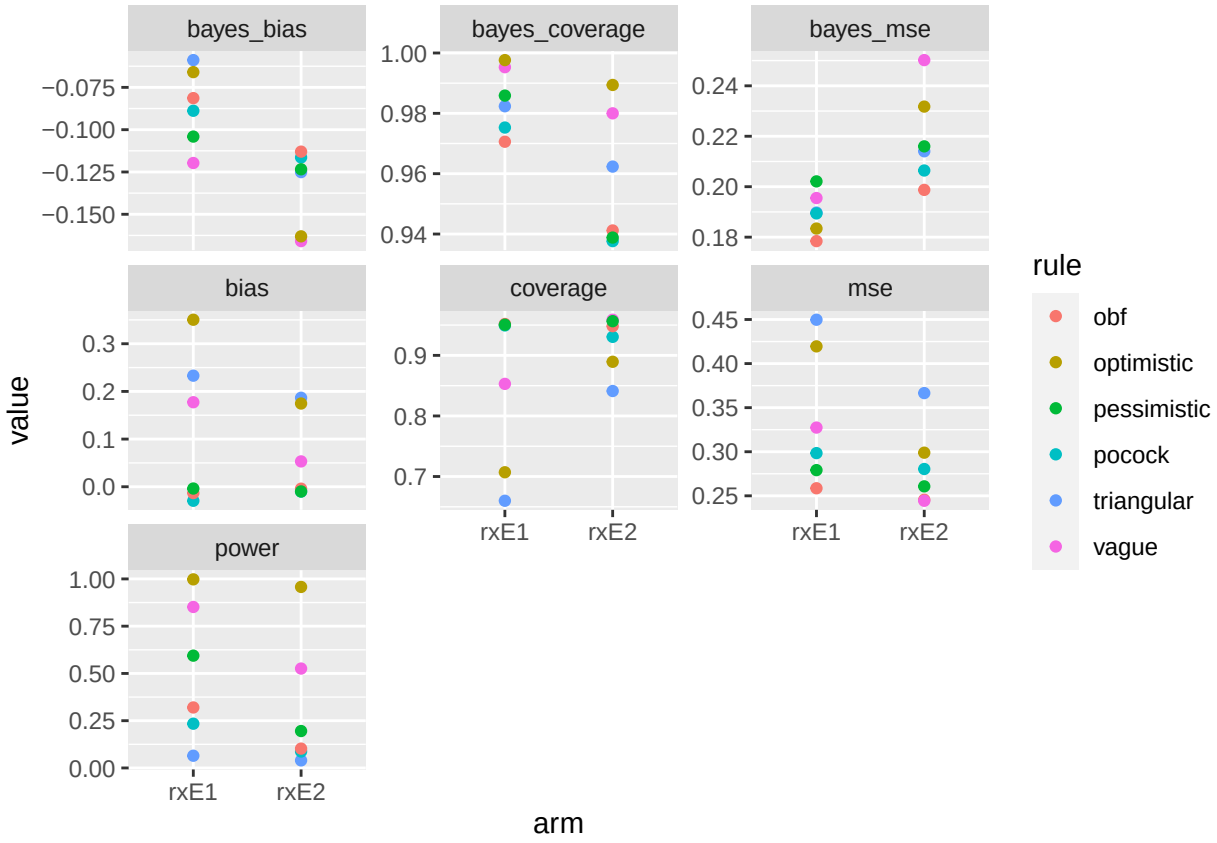


Positive but one arm < MCID

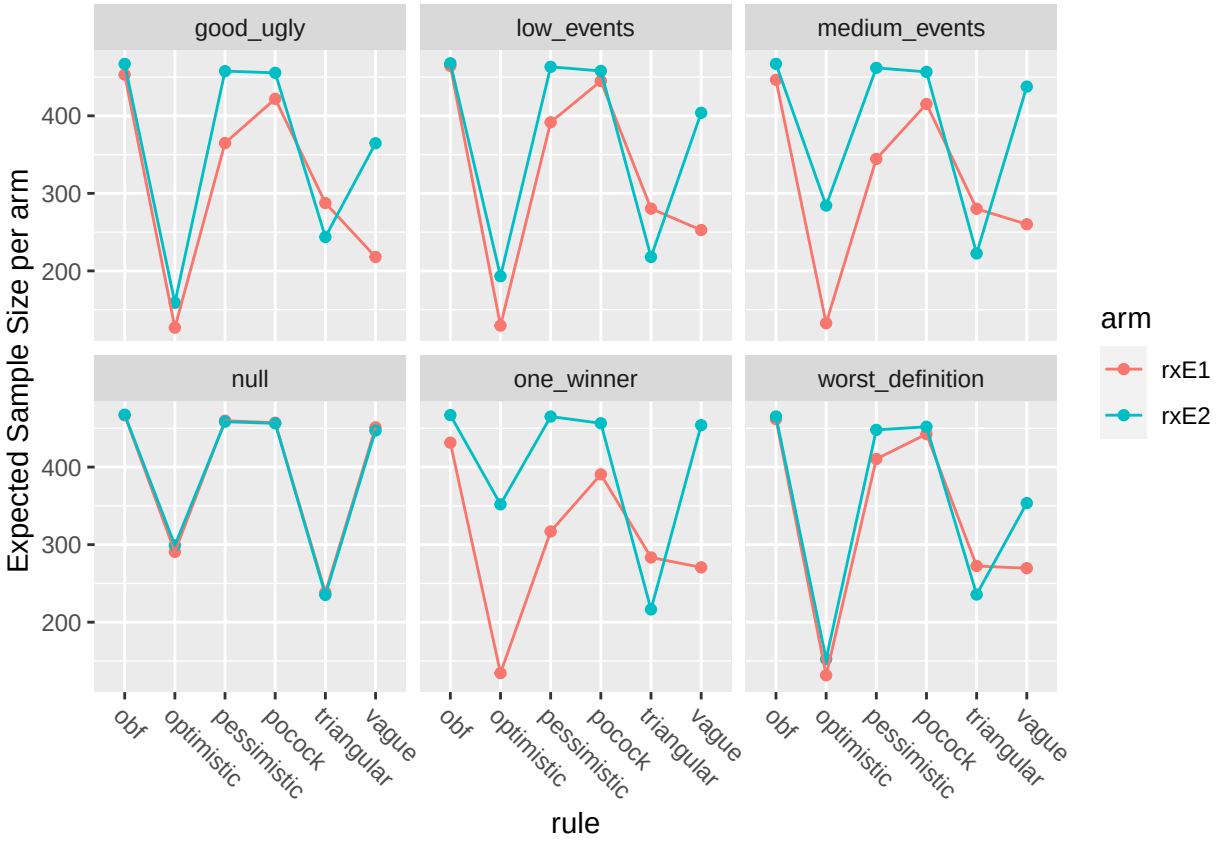
Good-Ugly HR= 0.6 and 0.85, control rate =0.9



Worst Case HR= 0.7 and 0.85, control rate =0.9



Expected Sample Size



Conclusions

The classical group sequential designs all suceed in stricly controlling the type 1 error rates to 1-sided 2.5% rate, under teh null. The Bayesian *rules* (as opposed to final inference) with a vague prior go up to 10%. This might still be acceptable, particularly for Tactic-R, but maybe not with TACTIC-e that wishes to acheive licencing potentially. The bayesian rule with an optimistic prior has an unncaptably high type1 error rate.

OTherwise, the Bayesian rules generaly achieve better power for alternativ hypotheses.

The coverage is broaldy acceptable across most choice of ryles and inference.

For biases, the Bayesian inference has small-modest bias for all the rules. THe frequentist inference give small-modest bias for teh freuqentist rules, but does poorly for Bayesian rules.

Sample size shows similar patterns across different scenarios

- optimistic stop early often
- triangular stops early but slight bigger ESS compared to optimistic
- O'Brien-Flemming, Pessimistic, Pocock, recruit the maximum size generally
- Vague stops early if there is a large treatment effect, but otherwise recruits the maximum size generally.

If I had to make a recommendation I would choose to use teh Vague prior decision rule, and perfom Bayesian inference as teh final analysis.

This simulation has been based around there being a rigid group sequential design with 3 fixed possibilities of sample size per arm. There is plausibly more flexibility in reality. I coudl investigate formulate rules to choose

sample sizes at interims. But it is not clear if there is a theoretical argument to stick to group sequential designs.

Are there any further simulations that shoudl be done at this point.

Could we anticipate what further simulations might be needed at the time of interims, and try to pre-prepare reports, and/or code.

There is a wider questino of the overall fate of this study. Both are being expanded to INdia, Mexico, Brazil. Tactic-R in India is a two-armed study with one of the active arms not being assigned. If the UK trial does not recruit, substantively then the chocie of decisions to be made at interim analyses is reduced; at the time of writing (Sep 2020) there is a second wave possibly startign to happen.

Risk that the code is at fault in some way, given the complexity of this simulation. Please do comment if anythign looks at odds with your knowledge. Happy to look closer wherever suggested.