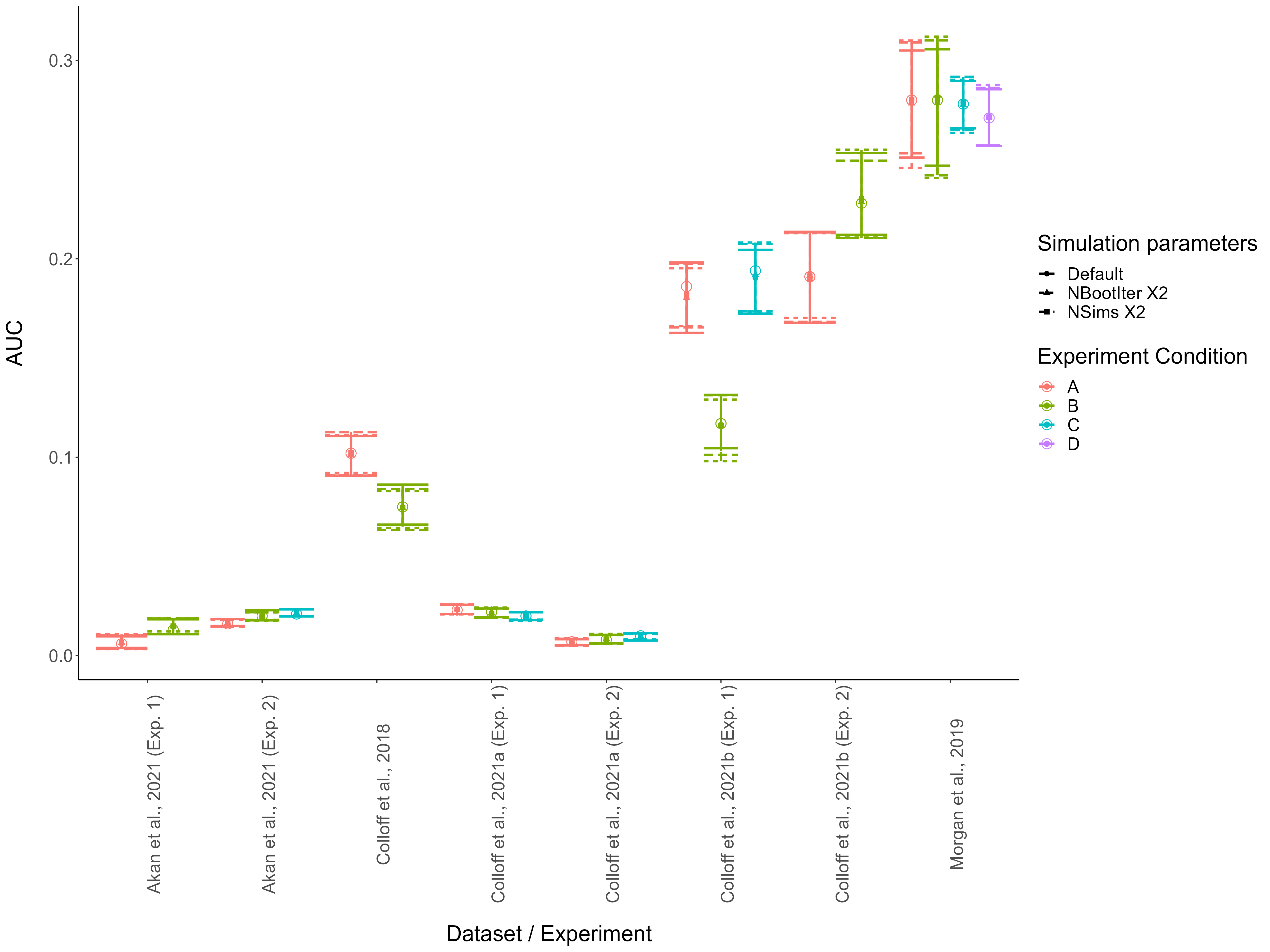
powe(R)OC Testing Results

Eric Mah

17/09/2021

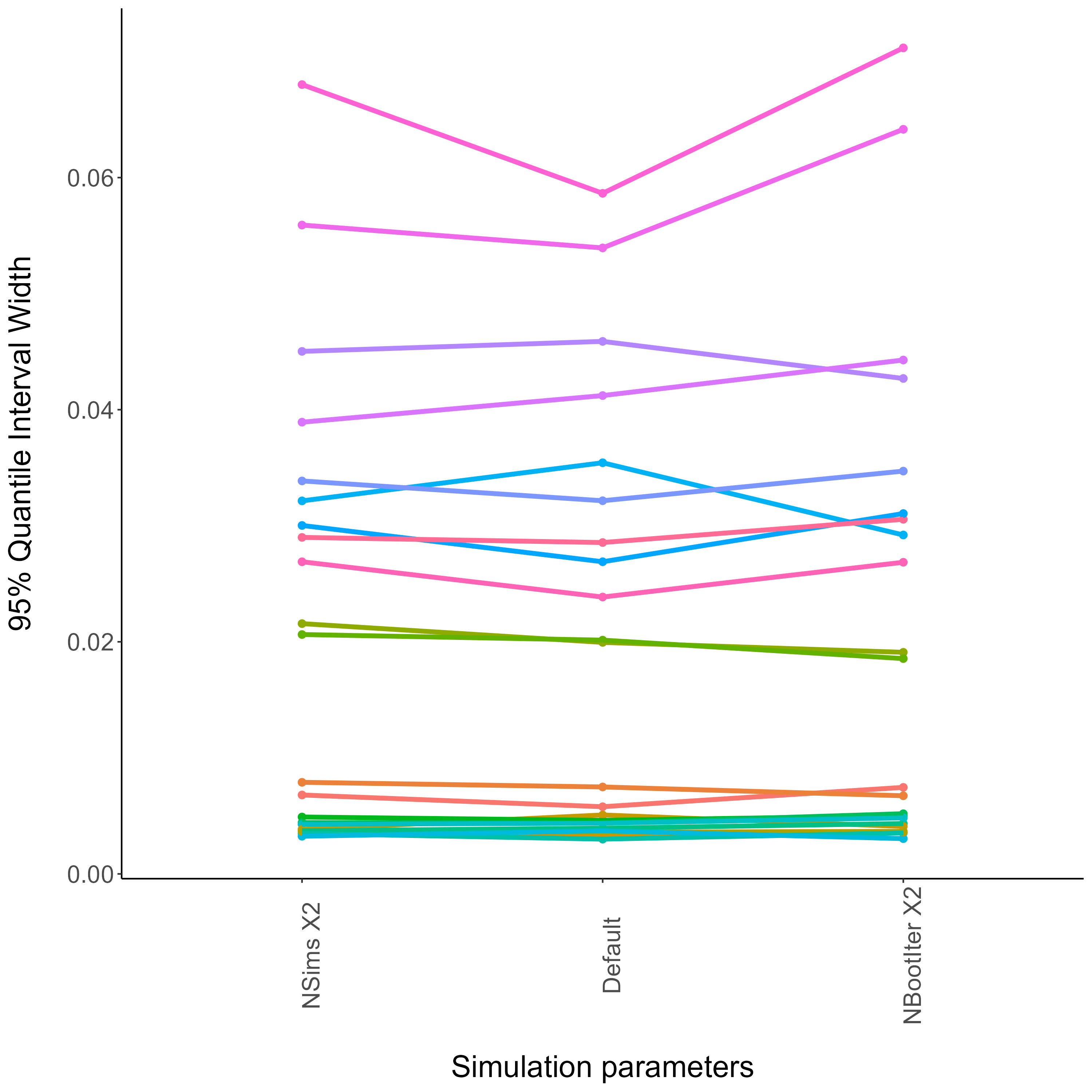
This document details several tests of simulation validity/performance conducted using real datasets. Data/articles used for testing can all be found in <https://github.com/E-Y-M/poweROC/tree/main/Dataset%20testing%20and%20reports>, and were obtained from the Open Science Framework. If you have ROC data (along with analysis parameters) you are willing to share for the purposes of simulation testing, feel free to email me at [ericmah@uvic.ca](mailto:ericmah@uvic.ca). Issues/comments on the app or simulation testing results can be posted on GitHub at <https://github.com/E-Y-M/poweROC/issues>.

## AUC Recovery

At a basic level, simulation validity depends on the ability of the simulations to recover AUC values close to those in the original dataset. The figure below depicts original AUC estimates from various papers with open data (5 papers, 8 experiments, 22 ROC curves computed using the same N’s/pAUC cutoffs in the original papers), along with simulated estimates and intervals: 

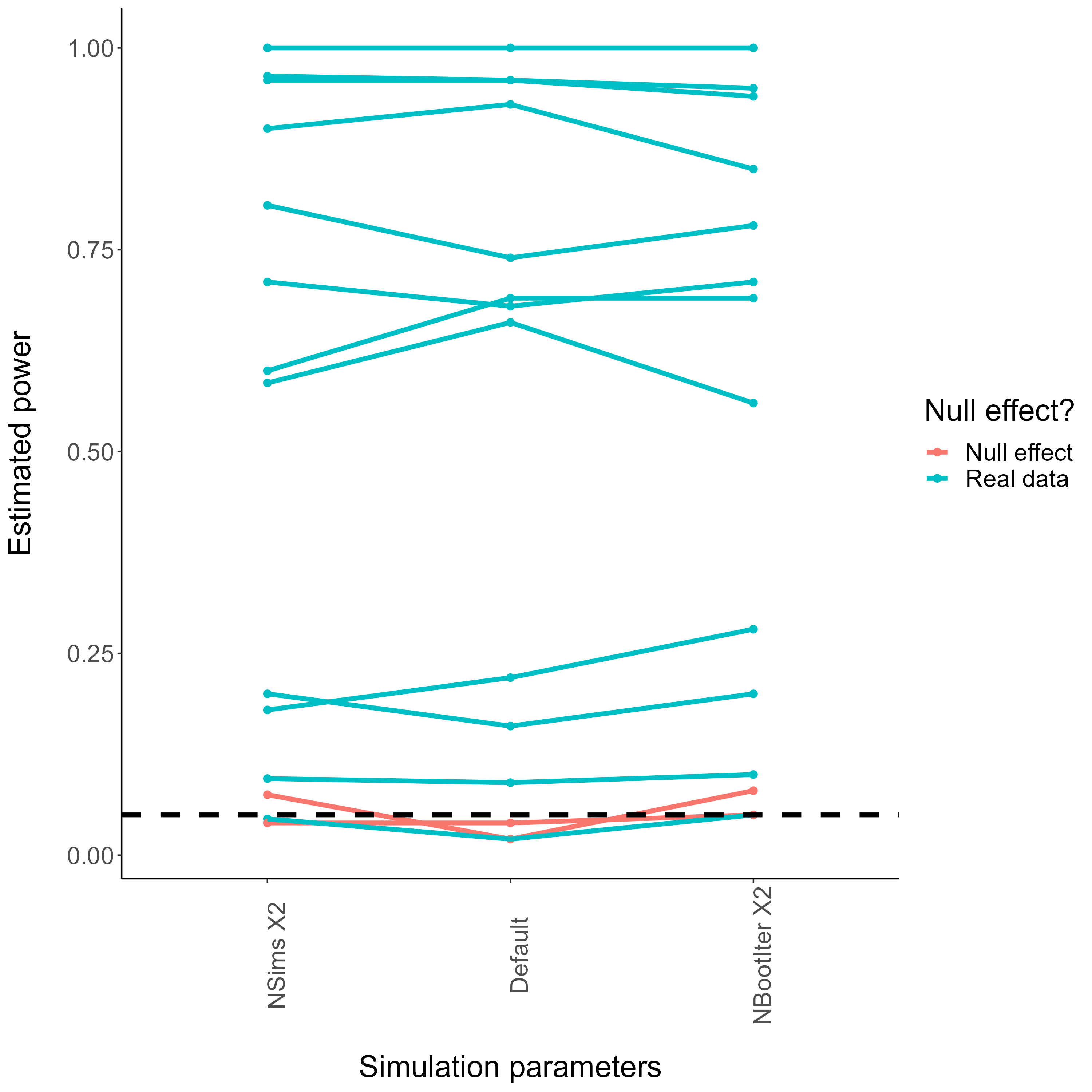
Testing the ability of the simulation to recover AUC values from experiments. Open circles represent original AUC values, all other points represent simulation estimates under various conditions (“NSims” = Number of simulated datasets per effect size/N, “NBootIter” = Number of bootstrap iterations per AUC comparison). Error bars = 95% quantiles on the mean estimated AUC for the simulations. Overall, simulations demonstrate excellent ability to recover original AUC values, even under default settings.

## Simulation precision under different conditions

Still, the question remains as to whether increasing the number of simulations or bootstrap iterations increases power. The figure below shows the width of the 95% quantile intervals for the AUC estimates above, as a function of the simulation conditions. 

Based on these simulations, it does not seem that increasing the number of simulations or bootstrap iterations necessarily or substantially increases the precision of the simulation beyond the default settings, suggesting that the default settings will result in reasonable estimates.

## Power estimates under different conditions

Finally, it is not clear whether the default simulation settings result in the most accurate power estimates. I simulated power for 13 ROC comparisons from the papers above along with two simulation runs of a dataset with a prespecified null effect (to compare with the normative Type I Error Rate of .05), all under the three different simulation conditions. These power estimates are plotted below: 

Power estimates differed slightly across the different simulation conditions, but no clear patterns emerged. In these examples, the maximum range of estimated power was .10. Importantly, power estimates in the two null effect simulations were close to the nominal Type I Error Rate of .05 (though the non-default settings resulted in slightly higher estimates). In light of these results, I recommend that users: a) Use the default simulation parameters, b) Power slightly higher than their target power (e.g., + .05-.10), and c) Conduct a couple simulation runs.