We can use this document to start jotting down ideas, questions, and comments to help start organizing some simulations to evaluate the different models across various scenarios of interest. Feel free to add anything or new sections, etc.

* Outcomes to potentially consider
  + Longitudinal proportional odds model
    - Have to consider prior specification(s)
  + Proportional odds model at day X
    - Likely want to explore a few days since Day 14 might have no difference between groups and would miss the potential benefit of fewer days with symptoms
  + AUC of the categorical outcomes
    - This assumes the 1-7 ordinal categories represent a good setting, but may be useful to examine potential power and could then use linear regression models
  + Survival as time-to-resolution of symptoms
    - Potentially need to make some simplifying assumptions
  + Naïve comparison of average number of days to resolution
* Missing data
  + At least 3 cases: (1) no missing, (2) some days missing but still observe outcome, (3) at some point they stop and never come back
  + Need to consider imputation strategies for the methods above based on the missingness
  + Consider scenario with similar patterns to our TREAT NOW data, also different patterns
* Simulation scenarios
  + We can use the TREAT NOW power calculations for one scenario
  + Will need to simulate null scenarios for type I error rates
  + Can also estimate potential transition probabilities based on TREAT NOW data to get an idea for other scenarios or what might be reasonable for some duration(s)

# Paper 1 – Calibration and Performance of Upstrapping for Futility

This paper will examine how one might want to calibrate for upstrapping and its performance relative to more standard group sequential designs.

1. Explore 4 different sample sizes (total maximum N=40, 160, 600, and 2000) across 4 different powered scenarios (5% [type I error], 50% [underpowered], 80% [adequately powered], 95% [overpowered]) and 4 total looks of the data (¼, ½, ¾, 1 [i.e., trial complete])
   1. If we assume a 60% response rate in one group, the above combinations for the overall data would correspond to response rates in the other group of:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Size** | ***Power for Simulation Setting*** | | | |
| **5%** | **50%** | **80%** | **95%** |
| N=40 | 60% | 29.2% | 18.0% | 9.5% |
| N=160 | 60% | 44.5% | 38.0% | 32.1% |
| N=600 | 60% | 52.1% | 48.6% | 45.4% |
| N=2000 | 60% | 55.7% | 53.8% | 52.0% |

* 1. Could also calculate the exact proportion from the table above with power.prop.test(n=Ngroup, p2=0.6, power=P, sig.level=0.05)$p1
  2. These sample sizes might be framed as ranging from phase IIa to phase III trials
  3. Can analyze with a Fisher’s exact/chi-squared test (hopefully will speed things up):
     1. Can automate choice with a few variations of code. One that works well is:
        1. catk <- chisq.test( tab )
        2. if( sum(catk$expected < 5) > 0 ){ catk <- fisher.test(tab) }
        3. Note, may need to create tab with a factor for outcome to ensure it is 2x2

1. Questions we wish to answer with these simulations (i.e., sections of the paper):
   1. How does upstrapping change with different scenarios:
      1. Within a power setting, how does the shading of the heat map change as sample size increases?
      2. When comparing null versus alternative, how does the shading of the heat map change?
   2. Assuming the adequately powered (i.e., 80%) scenario, does it matter how we calibrate the design? At each fraction of data (¼, ½, ¾), what is the combination of thresholds to maintain a desired type I error rate (e.g., 5%) while maximizing power if we:
      1. Fix the p-value used (e.g., p<0.05) and calibrate the proportion of upstrapped samples meeting the criteria *(note, only need the null scenario for calibration and then we can see how power performs)*
      2. Fix the proportion of upstrapped samples meeting a criteria (e.g., >80%), but calibrate the p-value needed *(note, only need the null scenario for calibration and then we can see how power performs)*
      3. Allow *both* the p-value and proportion to vary:
         1. Performance across all admissible combinations
         2. Performance if we take the designs nearest to the desired type I error threshold while maximizing power
   3. What is the performance of upstrapping across the different scenarios when we use different methods and calibration for futility?
      1. Methods to compare: the 4(?) calibration-selected methods above, one or two group sequential designs, the fixed sample design with no interim monitoring
      2. Measures to summarize:
         1. Power/type I error (depending on scenario)
         2. Proportion of trials that terminate early for futility
         3. Expected sample size across simulations
         4. Proportion of samples using Fisher’s versus chi-squared test for observed data (probably would be a brief comment in the results, but not much else)