colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Upstrapping To Determine Futility: Nonparametrically Predicting Future Outcomes From Past Data

Jess Wild MS, Alexander Kaizer PhD (University of Colorado Anschutz Medical Campus)

Introduction

Clinical trials often involve futility monitoring to reassess the chances of a successful trial based on the data collected at that point. Since the trial is ongoing, this available dataset includes only a fraction of the planned sample size, so estimating trial futility is challenging.

This simulation study evaluated the use of the upstrap algorithm for futility monitoring under a variety of simulation settings. Specifically, simulation scenarios varied by interim stopping point, power or type I error rate, and sample size.

Methods

Upstrapping is a nonparametric method that resamples the available data to supplement data already collected until a new dataset is generated that matches the desired total sample size of the trial. Resampling is done within each treatment group to preserve a 1:1 allocation ratio. The steps for applying upstrapping to an interim dataset are:

1. Resample *with replacement* from the observed data up to the expected total enrollment.
2. Calculate the p-value or posterior probability for the upstrapped "complete" dataset.
3. Repeat a large number of times (e.g., 1000).
4. Calculate the proportion of upstrapped p-values that meet some threshold (e.g., $p < 0.05$, posterior probability > 0.95).

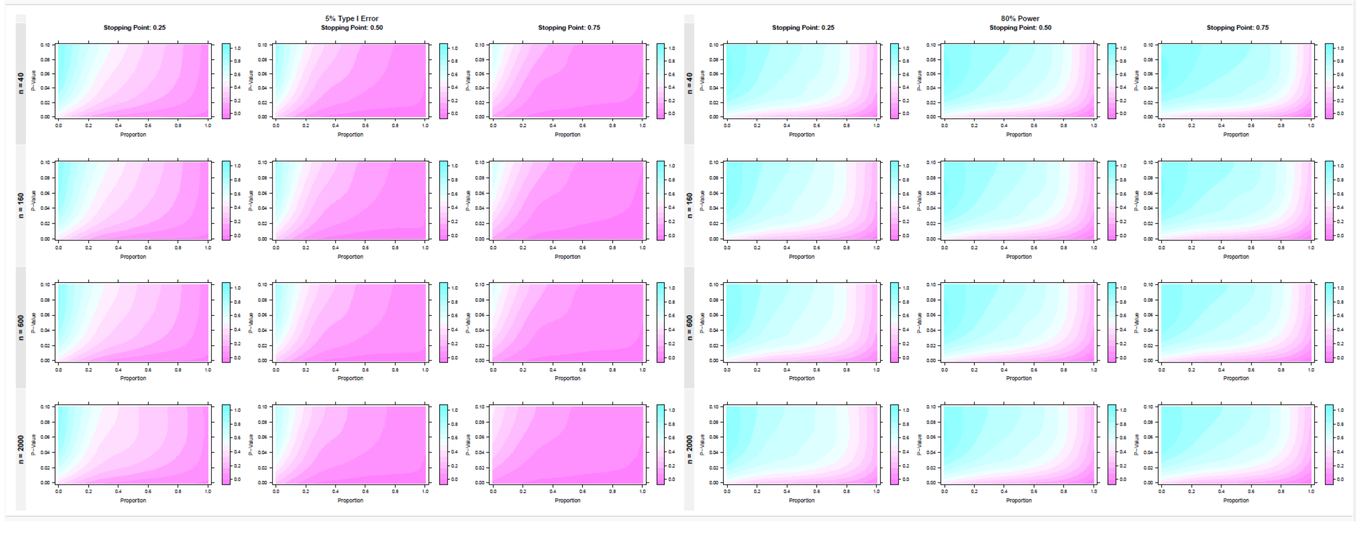

Figure 1: Method Validation Results *Results reported as heatmaps showing the probability of meeting the defined p-value and proportion combination (blue representing more likely to meet the criteria, pink representing less likely) for various p-value (y axis) and proportion (x axis) threshold combinations. The null (5% type I error, left side) and alternative (80% rower, right side) scenarios are presented with subplots faceted by information fraction at the interim look (0.25, 0.50, 0.75 from left to right) and maximum trial sample size (40, 160, 600, 2000 from top to bottom).*

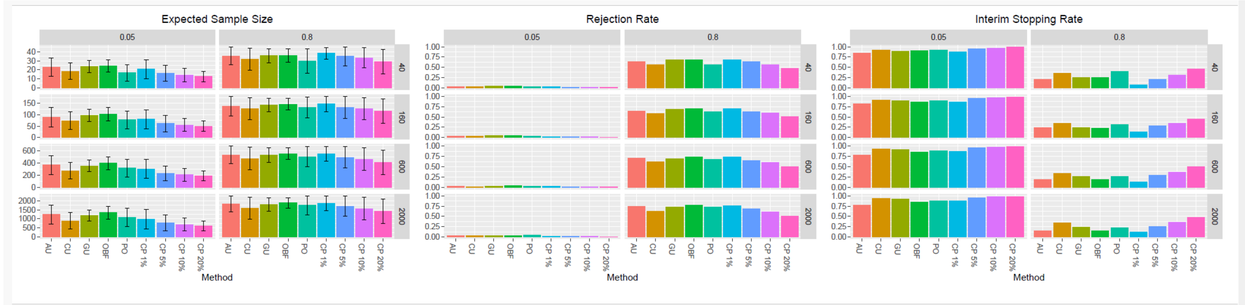

Figure 2: Main Analysis Results *The left panel shows mean expected sample size (y axis) reported with error bars representing ±1 SD for each interim monitoring method (x axis). Graphing scale is relative to total sample size. The middle panel shows rejection rate results, where rejection rate is defined as the proportion of simulated trials that reached trial completion and then rejected the null hypothesis. Rejection rate (y axis) is reported for each interim monitoring method (x axis). The right panel shows interim stopping rate results with interim stopping rate defined as the proportion of simulated trials that stopped early (y axis) reported for each interim monitoring method (x axis). Subplots are faceted by power/type I error rate (5% or 80% from left to right) and total sample size (40, 160, 600, 2000 from top to bottom).*

Futility Monitoring Methods:

AU	***Arbitrary Upstrap:*** apply proportion threshold of 80% and p-value threshold of 5%
CU	***Calibrated Upstrap:*** calibrate proportion and p-value thresholds to optimize power and type I error rate
GU	***Group Sequential Upstrap:*** calibrate proportion threshold to optimize power and type I error rate and apply p-value threshold derived from group sequential methods
OBF PO	***O'Brien-Fleming and Pocock:*** traditional group sequential interim monitoring methods used as a comparator for the upstrap
CP	***Conditional Power:*** alternative comparison method, extrapolates overall power to detect an effect conditional on the current data using a specific probability threshold (1%, 5%, 10%, 20%)

Conclusions

- Many futility monitoring methods exist but they typically rely on limiting parametric assumptions about the data, while upstrapping is a nonparametric method not requiring such assumptions
- Upstrapping performs generally well for futility monitoring.
- For a given sample size, the likelihood of finding a significant result was comparable across interim stopping points.
- Similarly, the likelihood remained roughly equivalent across sample sizes at the same interim stopping point.
- Expected sample size, rejection rate, and interim stopping rate results all show that, while trade offs exist, upstrapping performs favorably when compared with traditional methods

Colorado School of Public Health | 13001 East 17th Place, 3rd Floor, Mail Stop B119 | Aurora, CO 80045
coloradosph.cuanschutz.edu