Clinical Trials
XX(X):1–??
©The Author(s) 2023
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Upstrapping to Determine Futility: Predicting Future Outcomes from Past Data

Jessica L. Wild, MS[1] and Alexander M. Kaizer. PhD[1]

Abstract

Background: Clinical trials often involve some form of interim monitoring to determine futility and/or efficacy before planned trial completion. While many commonly used parametric options for interim monitoring exist, nonparametric based interim monitoring methods are also needed to account for more complex trial designs and analyses. The upstrap is one recently proposed nonparametric method that may be applied for interim monitoring.

Methods: Upstrapping involves repeatedly resampling existing interim data with replacement and determining the likelihood of trial success based on comparison to preselected p-value and proportion threshold criteria. To evaluate the potential for upstrapping as a form of interim monitoring, we conducted a simulation study considering different sample sizes. We first compared trial rejection rates across a selection of threshold combinations to validate the upstrapping method. We then considered several different calibration strategies for these thresholds. Finally, we applied upstrapping methods to simulated clinical trial data, directly comparing their performance with more traditional alpha-spending interim monitoring methods.

Results: The method validation results showed that upstrapping can produce reasonable results across a variety of simulations settings. Although there are many potential approaches to calibration, our three proposed approaches had different strengths depending on the stopping rules used. For futility only interim monitoring, upstrapping performed similarly well across performance metrics compared to alpha-spending methods. For efficacy only interim monitoring, upstrapping methods produced promising results with regards to reductions in the expected sample size, but did inflate the type I error rate in null scenarios. For combined futility and efficacy monitoring, similar trends were seen with large reductions in the expected sample size but a trade-off in type I error control.

Conclusions: Upstrapping is a promising tool for performing interim monitoring, particularly for futility monitoring. When properly calibrated, upstrapping is found to be comparable to alpha-spending methods in accurately predicting trial futility but with a smaller expected sample size.

Keywords

Interim monitoring, futility monitoring, efficacy monitoring, upstrap, nonparametric, alpha-spending

[1] Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, Colorado

Introduction

Interim monitoring is an essential part of many clinical trial designs meant to increase efficiency by stopping trials that show particularly strong signs of either futility or efficacy before their planned endpoints. This allows time and resources to be directed towards interventions that show promising results early on in the process and away from interventions that are likely to yield null results at the end of the trial. Interim monitoring is commonly used in trials of many different designs across all phases of the research process, with multiple stopping points allowing the potential to stop a trial early at several points before full data collection is complete. Depending on the circumstances of the trial, interim monitoring may be used to stop only for futility, only for efficacy, or a combination of the two. Interim monitoring should always be accounted for at the design stage of a trial to avoid an inflated type I error rate or a reduction in statistical power (1; 2; 3).

There are many methods available to perform interim monitoring, including group sequential designs and alpha-spending functions including O'Brien-Fleming, Peto, and Pocock boundary methods (1; 2; 3; 4; 5; 6; 7). Group sequential designs generate p-value boundaries for futility and/or efficacy to be applied at the planned stopping points and at conclusion of the trial (7). These designs also account for multiple interim stopping points to maintain the desired trial operating characteristics, such as power and the type I error rate. While group sequential methods are commonly used, they each involve certain distributional assumptions which make them less suited to complex statistical models or data that violates these parametric assumptions.

This research project is motivated by the need for nonparametric methods for interim monitoring in the TREAT NOW clinical trial (NCT04372628). TREAT NOW was a phase 2b multi site clinical trial comparing the clinical use of lopinavir/ritonavir to placebo in an outpatient context for the treatment of COVID-19 (8; 9). The primary outcome used a longitudinal ordinal proportional odds Bayesian logistic regression model. Because the planned analysis was complex, traditional interim monitoring methods were not available and nonparametric approaches were considered instead.

One recently proposed nonparametric framework that could be applied in the context of interim monitoring is the *upstrap* (10). Upstrapping relies on repeatedly resampling incomplete data to impute future observations. In a clinical trial context, this could be used to predict chances of trial success. The method has been implemented to perform interim monitoring in clinical trials (11). However, there has not yet been a thorough review of the upstrap method's performance or validity when used for interim monitoring. In this paper we use simulation studies to evaluate the general performance of the upstrap algorithm when used for interim monitoring. We first introduce the concept of the upstrap as applied to interim monitoring in clinical trials and potential calibration strategies to identify stopping rules. A simulation study assuming a binary outcome for simplicity is then used to elucidate the properties of the upstrap and compare to alpha-spending methods for interim monitoring. We conclude with a discussion of this newer approach to interim monitoring and settings where the proposed calibrations may be appropriate.

Methods

Upstrap Algorithm

Upstrapping is a generic approach that resamples the available data (with replacement) to supplement data already collected until a new dataset is generated that matches a desired total sample size. In the context of a clinical trial,

Corresponding author:

Jessica Wild

Email: jessica.wild@cuanschutz.edu

this represents the planned maximum sample size to enroll assuming no early termination at an interim analysis. The resampling is done within each treatment group to preserve the desired allocation ratio (e.g., 1:1 between study arms). The steps for applying upstrapping to an interim analysis dataset are:

(i) Resample with replacement from the observed data up to the expected total enrollment.

(ii) Calculate the p-value for the upstrapped “complete” dataset.

(iii) Repeat a large number of times (e.g., $N_{U_p} = 1000$).

(iv) Calculate the proportion of upstrapped p-values that meet a set p-value threshold (e.g., $p < 0.05$).

(v) Compare the calculated proportion to the set proportion threshold (e.g., $P > 0.80$)

Notably, this means that set p-value and proportion thresholds must be determined a priori. At any interim stage, the upstrap could be used to estimate the probability that the trial will be ”successful” based on the given thresholds, with a decision made with regards to potentially stopping the trial for futility, efficacy, or both.

Simulation Settings

Each simulation setting included a binary outcome measured once per subject with subjects assigned to either treatment or control. A variety of simulation settings were considered based on three varying parameters: sample size, power, and interim analysis stopping point. Sample sizes considered include a wide range to represent trials at different research stages, including per group total sample sizes of 20, 80, 300, and 1000. Power includes cases representing the standard 80% power alternative scenario and 5% type I error rate null scenario, as well as underpowered (50% power) and overpowered (95% power) scenarios to reflect real world uncertainty. Interim stopping points were based on proportion of subjects enrolled (0.25, 0.50, or 0.75) and are considered to be sequential within a trial. For each setting we considered p-value thresholds between 0-0.1 and proportion thresholds from 0-1 when applying the upstrapping algorithm. Chi-squared or Fisher’s exact test, depending on test assumptions, were used to estimate the treatment effect in every upstrapped dataset as well as the full sample dataset. For full sample analyses a p-value of less than 0.05 was considered to be significant (except for group sequential design applications which adjust the full sample p-value to control type I error). Upstrapping was performed using $N_{U_p} = 1000$ upstrapped datasets at each interim analysis. Each simulation setting was repeated 1000 times using R v4.0.2 (Vienna, Austria).

Method Validation

The first research aim was to validate use of the upstrap method for interim monitoring. This involved using simulation results to evaluate the method’s chances of stopping based on a grid of p-value and proportion thresholds (defined as a 20 x 20 grid with p-values between 0-0.1 and proportions between 0-1). This grid of threshold values was applied to each simulation setting, providing results across all sample size, power, and stopping point combinations. To simplify the validation, efficacy only interim monitoring is presented, but a similar analysis could also be done for futility only or combined monitoring. Essentially, the goal of this research aim is to compare how often the upstrap method determines a trial will stop based on a variety of threshold values, then evaluate how this proportion grid changes between simulation settings.

Method Calibration

Since the upstrap method relies on two key threshold values, p-value and proportion, we must also consider how to properly calibrate these threshold values based on the simulation results. Using the grid of potential p-value and proportion threshold values discussed in the Method

Validation section, several different calibration approaches were considered and the general approach to their calibration is as follows:

Arbitrary Calibration (AU) assumes the desired alpha-level for the p-value threshold with futility stopping assuming the proportion of upstrapped samples less than the desired type I error rate and efficacy stopping assuming the proportion greater than the desired power. We chose $p < 0.05$ and $P < 0.05$ for the p-value and proportion threshold criteria for futility monitoring, and inversely $p < 0.05$ and $P > 0.80$ for efficacy monitoring. ***Variable Calibration (CU)*** uses a pre-specified grid of potential p-value and proportion thresholds and then identifies the p-value and proportion threshold combination needed for futility or efficacy monitoring to achieve a desired level of type I error rate or power. ***Group Sequential Inspired Calibration (GU)*** uses the p-values from the alpha-spending O'Brien-Fleming boundary for the p-value threshold and searches for a corresponding proportion to achieve a desired level of type I error or power.

For each calibration approach, the rejection rates were calculated (i.e., type I error rate for null scenarios, power for alternative scenarios). Threshold values were then chosen based on attempting to optimize these characteristics, considering all possible threshold combinations without preference. For efficacy monitoring, only threshold combinations producing a type I error rate of at most 5% were considered, then the combination with maximum power was selected from these candidates. For futility monitoring, only threshold combinations producing a type II error rate of at most 20% were considered, then the combination which maximized power was selected from these candidates. This process was done separately for each sample size and stopping point within the CU and GU calibration approaches.

Method Application

To evaluate the performance of the upstrap algorithm we applied both upstrapping and group sequential methods to the simulation results to perform interim monitoring. We considered the three different calibration approach described previously, as well as alpha-spending functions with O'Brien-Fleming (OBF) and Pocock (PO) style boundaries. Since there are multiple ways to structure interim monitoring, we considered three interim monitoring types: futility only (FO), efficacy only (EO), and futility and efficacy combined (FE). Interim stopping points of 0.25, 0.5, and 0.75 were used sequentially within each simulated trial.

Each simulation setting is summarized by the mean and standard deviation of expected sample size across simulation replications, the proportion of trials that stopped early, and the proportion of trials that rejected the null hypothesis. The FE monitoring results also summarize the proportions stopping specifically for futility or efficacy separately. For comparison, the proportion of trials that rejected the null hypothesis with a fixed sample (FS) design without interim monitoring was also calculated to serve as the reference point for each simulation setting.

Results

Performance of Upstrapping Across Sample Sizes and Information Fractions

Before applying a specific calibration strategy (e.g., AU, CU, or GU), we first evaluated the general trends across sample sizes and information fractions across a grid of p-values and proportion of upstrapped samples less than or equal to that p-value. This allows for a validation of the general concept of upstrapping, which we present for both the null and alternative scenario.

Figure 1 presents heatmaps representing the probability of stopping the trial for efficacy based on the grid of threshold combinations, with different plots showing variation due to

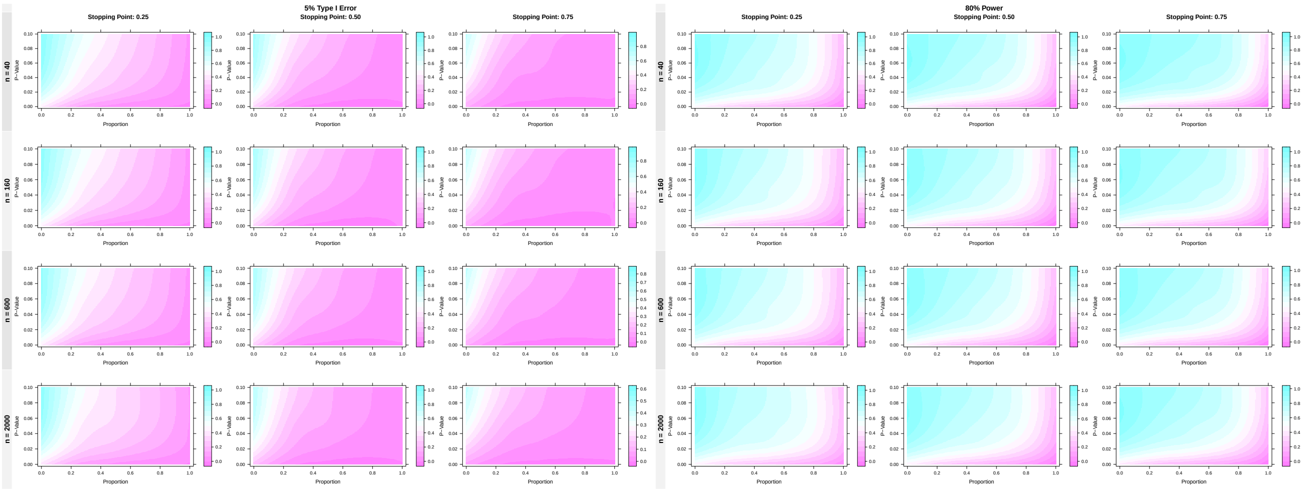

Figure 1. *Method Validation Results:* Results reported as heatmaps showing the probability of stopping a trial early for efficacy (blue representing more likely to stop, pink representing less likely to stop) for various p value (y axis) and proportion (x axis) threshold combinations. A separate plot is presented for the null (5% Type I Error, left side) and alternative (80% Power, right side) settings and subplots are faceted by interim stopping point (0.25, 0.50, 0.75 from left to right) and total sample size (40, 160, 600, 2000 from top to bottom).

sample size, power, and stopping point (i.e., the observed information fraction). Results indicate that across sample sizes and stopping points the upstrap is much more likely to stop early for efficacy across different combinations in the alternative (80% power) case rather than the null (5% power) case. These findings are encouraging and indicate that the method produces expected results when used to perform interim monitoring. Additionally, results show similar trends across stopping points within a given sample size, as well as between sample sizes at the same stopping point. Overall, the validation results provide evidence that upstrapping may be a reasonable approach for interim monitoring.

Sequential Monitoring Results

This section presents the results for the three proposed upstrap calibration approaches (AU, CU, and GU) and two alpha-spending functions (PO and OBF) when there are three interim looks at 0.25, 0.50, and 0.75 under both null and alternative scenarios. The results are presented separately for futility only (FO), efficacy only (EO), and futility or efficacy (FE) monitoring. Each interim stopping rule starts with the traditional alpha-spending results before introducing the results for the upstrap approaches.

Figures 2-4 present the expected sample size, rejection rates, and interim stopping rates, respectively. Tables 1 and 2 present the results for each interim monitoring strategy as the difference in rejection rate from a fixed sample (FS) design without interim monitoring and the ratio of the expected sample size to the fixed sample size for the null and alternative scenarios, respectively. We additionally describe the general performance of the upstrap calibrations if only one interim look is conducted at 0.25, 0.5 or 0.75. The specific numeric summaries are presented in the Supplementary Materials in Tables S1-S14.

Futility Only (FO) Monitoring Results The OBF method has type I error rates within 0.4% of the FS design in the null scenario and reductions of approximately 2% to power in the alternative scenario. The ESS across all sample sizes is about 70% of the FS in the null when we expect to stop for futility, and around 95% of the FS in the alternative when stopping for futility is suboptimal. The PO method has similar type I error rates with even smaller ESS (around 50% of FS), but has a 10% decrease in power relative to the FS design across all sample sizes.

The CU upstrap calibration performs most similarly to the PO, albeit with a greater reduction in power at larger sample sizes of up to 18.5%. The AU and GU upstrap calibrations have type I errors within 1% of the FS design while needing only 58-62% of the ESS. However, the AU power is 4.8-7.1% lower with an ESS of 86-91% of FS, and the GU power increases from 1.9% to 10.1% lower as the sample size increases.

The OBF is the most balanced with respect to type I error and power trade-offs, but the savings in ESS are minimal relative to the reductions with AU or GU upstrap calibrations for a futility only monitoring design. If decreasing the expected sample size is an important consideration, AU or GU may represent an acceptable trade-off considering it is about 10% lower than OBF in all null scenarios. Additionally, the AU requires no prior calibration, resulting in a low barrier to implementation.

Efficacy Only (EO) Monitoring Results The OBF method has type I error rates within 0.8% of the FS design in the null scenario and power within 1.7% for alternative scenarios. In the null scenario, the ESS of the OBF is 1.00, indicating it never stops early for efficacy when there is no effect. For the alternative scenario, the ESS is 83% to 90% of the FS design. The PO method has similar performance in the null scenario, but has 6.8% to 10.8% lower power across the different sample sizes with ESS being 73% to 85% of the FS design. Generally, as the sample size of the scenarios increase, the relative proportion of the ESS to a FS design decreases.

The CU calibration has type I error rates that are 5.9-9.4% higher than the FS design, with the ESS in the null scenario being approximately 94% of the FS. Meanwhile, the power with the CU calibration is 5.0-8.6% higher with an ESS of 60-70% of the FS design. The GU calibration performs similarly, but is more consistent with type I error rates being 7.0-7.8% higher across all sample sizes. The AU calibration has an elevated type I error rate of 16.2-18.7% in the null scenario, while also having power that is 6.5-12.0% higher in the alternative scenario. The ESS of the AU calibration in the alternative is 51-55% of the FS, indicating a substantial savings in sample size.

The OBF achieves the most similar power and type I error rates to the FS design, but does so at the expense of less savings in terms of ESS. In contrast, the CU and GU upstrap calibrations result in increases to power with larger reductions in the alternative scenario ESS compared to alpha-spending approaches, but do so at the expense of an inflation to the type I error rate of up to 9.4%. The AU calibration when applied to only efficacy monitoring has the highest inflation to the type I error rate at up to 18.7%, which is unlikely to be acceptable for many trial settings.

Futility and Efficacy (FE) Monitoring Results The OBF method has type I error rates within 0.4% and power that is 0.7-2.2% lower compared to the FS design when both futility and efficacy interim monitoring are used. In the null scenario, the ESS is 68-71% of the FS design, whereas in the alternative it is 79-85%. The PO method has type I error rates within 0.9%, but power that is 13.1-18.7% lower. The ESS for PO in the null scenario is 40-50% and for the alternative scenario is 56-60% of the FS design.

The GU calibration has type I error rates that are 6.1-7.2% higher with an ESS of 55-58% that of the FS, whereas the power ranges from 4.3% higher to 4.9% lower as the sample sizes increase across scenarios with 52-58% of the ESS of the FS. The CU has lower inflation of the type I error rate at 3.9-6.0% with a samller ESS of 41-45%, but the power ranges from 5.1% to 14.1% lower compared to the FS design as the sample sizes of the scenarios increase. The AU calibration has type I error rates that are 15.4-18.1% higher than the FS design, with only a maximum increase in power of 6.6% in the smallest sample size scenario with n=40.

Again, the OBF method is fairly close to the observed type I error rate and power of the FS design, but has less improvement in the ESS compared to the upstrapping approaches when evaluating for both futility and efficacy. The GU method has the best balance between null and alternative scenarios, with some inflation in the type I error rates but a much smaller ESS in both settings than the OBF.

Results for Under- and Over-Powered Scenarios Simulations were also conducted to evaluate the performance of the interim monitoring methods if an underpowered (50% power) or overpowered (95% power) scenario were encountered when the design was calibrated assuming an effect size corresponding to 80% power. Figures 2-4 show these scenarios for each of the operating characteristics. The underpowered scenario performs between the previously described null and alternative scenarios. The overpowered scenario shows the greatest reductions in the ESS with efficacy monitoring and has high rejection rates for most methods except PO which has lower power relative to the other approaches.

Performance of Upstrap Calibrations with Only a Single Interim Look The upstrapping calibration performance when only conducting a single interim look differs from the design with three interim looks. Since only one interim look is conducted, the observed type I error rate and power of each scenario are closer to that of the FS design. The ESS is minimized with a single look at 0.5. At a single 0.25 look, the ESS is higher due to having a less information to upstrap the remaining 75% of the trial from. At a single 0.75 look, the ESS is also higher because the design only allows stopping after three-quarters of the study has been enrolled. The specific performance of these single interim looks are presented in the Supplementary Materials Tables S1-S14.

Discussion

There are numerous strategies for conducting interim monitoring within a clinical trial. In this paper we propose the use of the non-parametric upstrap as a potential interim monitoring strategy and evaluate its potential performance across a range of calibration strategies. While we focus on the use of the upstrap in a binary outcome setting to facilitate comparison to existing alpha-spending interim monitoring methods, the upstrap may provide a flexible framework for statistical models with complexities that have not been fully addressed by existing methods. For example, the longitudinal proportional odds logistic regression model used in the TREAT NOW clinical trial used an upstrapping approach (8; 9).

The validation of the general performance of the upstrap across information fractions and sample sizes suggested promising performance for interim monitoring. Figure 1 demonstrated that the null scenario heat maps have far lower probability of stopping a trial early for efficacy, whereas the alternative scenario heat maps had far greater probabilities of stopping early for efficacy across p-value and proportion of upstrapped samples below this p-value.

In order to operationalize the validation performance, we proposed three different calibration strategies to evaluate the trial operating characteristics of compared to existing Pocock (PO) and O'Brien-Fleming (OBF) alpha-spending approaches. The arbitrary (AU) approach was simple in that the desired type I and II error levels can be used. The calibrated (CU) approach considers the specific scenario (e.g., sample size, null and alternative response, information fraction). The groups-sequential (GU) approach to upstrapping used OBF p-value thresholds with the proportion of upstrapped samples calibrated. However, none of the approaches explicitly accounted for the repeated looks at the data and led to inflations of the type I error rates in scenarios with efficacy monitoring.

However, all upstrapping approaches do result in large reductions to the expected sample size (ESS) of the study relative to both PO and OBF alpha-spending approaches. Upstrapping approaches were generally more likely to stop a trial early compared to PO and OBF, as shown in

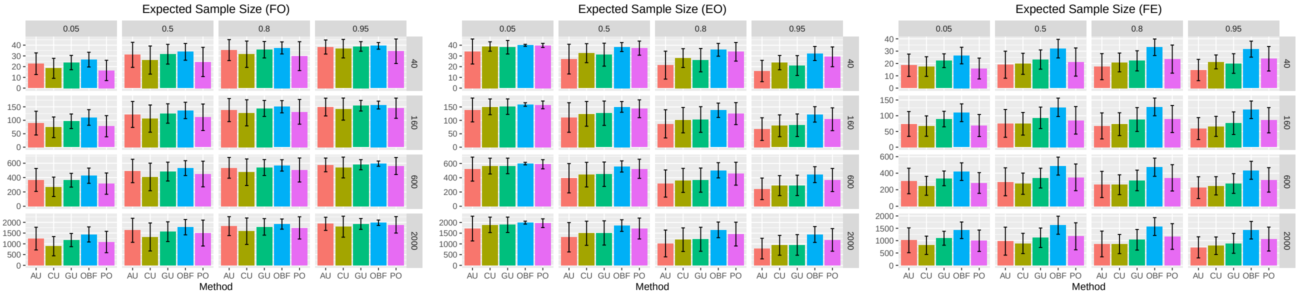

Figure 2. *Expected Sample Size Results:* Mean expected sample size (y axis) is reported with error bars representing ± 1 SD for each interim monitoring method (x axis). A separate plot is presented for each interim monitoring design (FO, EO, FE from left to right) and subplots are faceted by power (5%, 50%, 80%, 95% from left to right) and total sample size (40, 160, 600, 2000 from top to bottom). Graphing scale is relative to total sample size.

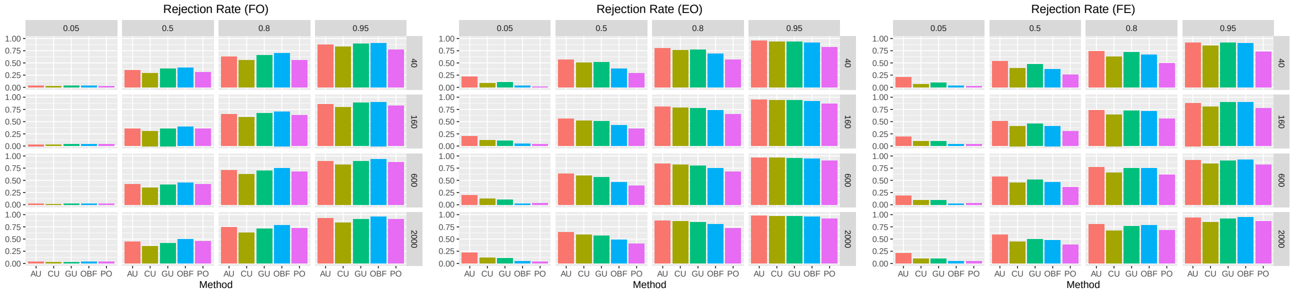

Figure 3. *Rejection Rate Results:* Here rejection rate is defined as the proportion of simulated trials that either stopped early for efficacy or reached trial completion and then rejected the null hypothesis. Rejection rate (y axis) is reported for each interim monitoring method (x axis). A separate plot is presented for each interim monitoring design (FO, EO, FE from left to right) and subplots are faceted by power (5%, 50%, 80%, 95% from left to right) and total sample size (40, 160, 600, 2000 from top to bottom).

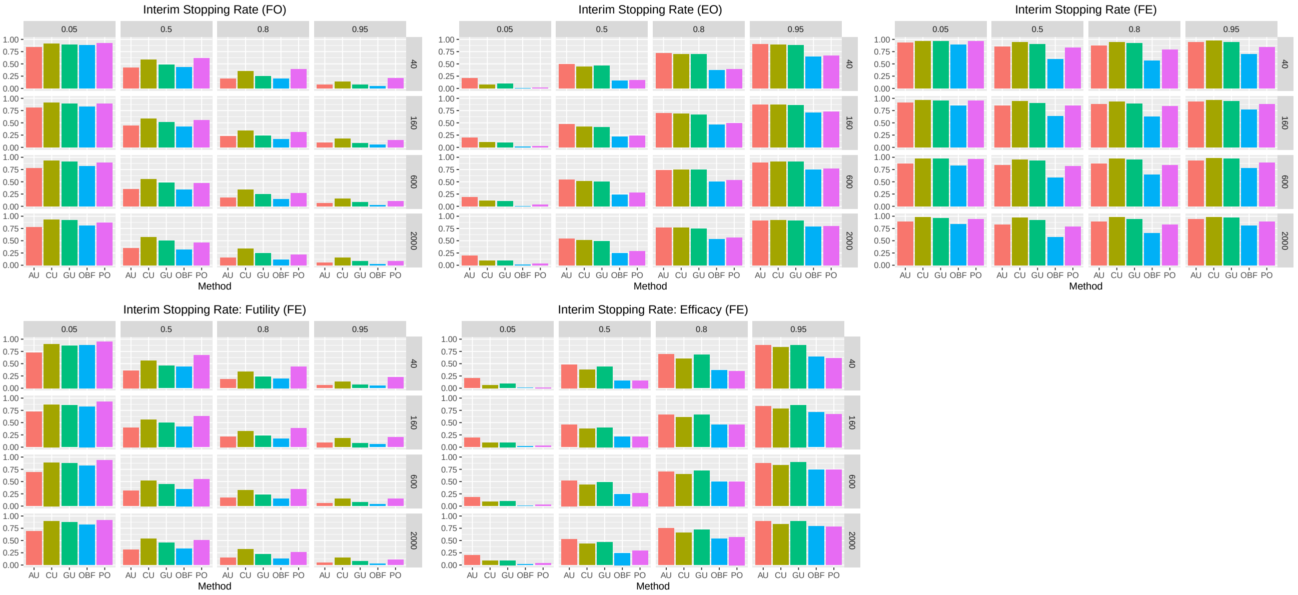

Figure 4. *Interim Stopping Rate Results:* Proportion of simulated trials that stopped early (y axis) is reported for each interim monitoring method (x axis). A separate plot is presented for each interim monitoring design (FO, EO, FE from left to right). For FE monitoring (where at each stopping point the option exists to stop for either futility or efficacy), we further include separate plots for proportion of simulated trials that stopped early specifically for futility (Futility (FE)) and specifically for efficacy (Efficacy (FE)). Subplots are faceted by power (5%, 50%, 80%, 95% from left to right) and total sample size (40, 160, 600, 2000 from top to bottom).

Figure 4. Across all interim monitoring settings of futility-only, efficacy-only, or both, the GU approach provided what may be acceptable trade-offs for in desired type I error rates or power considering the drastic reductions in ESS relative to OBF of up to 16% in null scenarios and 27.5% in alternative scenarios. The CU approach was

overly conservative for futility-only monitoring, but similar to GU when using efficacy monitoring. Interestingly, the AU approach performed well in futility-only monitoring and could be an easy-to-implement approach if lower power is tolerable.

This research is an exploratory study of whether upstrapping may be a potential approach for interim monitoring. Accordingly, many simplifying assumptions were made at both the simulation and modeling stages of our analysis. We considered simulation settings based only on sample size, interim stopping point, and power. All simulations assumed uniform subject accrual over time, and a constant treatment efficacy rate. It may also be worth considering more complicated modeling strategies, potentially with covariate information included, to reflect a wider range of possible study design settings.

Calibration and application results showed a clear trade off between power and type I error rate for the general upstrapping method, particularly when efficacy monitoring is needed. This is not unexpected, but is an important consideration when deciding on an interim monitoring method and choosing threshold values. In general, we considered several different approaches to threshold calibration and chose the best values based on power and type I error rate considerations. However, this process could easily be extended to consider a wider variety of approaches or a more granular grid of potential threshold values. Additional work is needed to develop calibration approaches for the upstrap which can better control the type I error rate while achieving the desired power. One possibility is to consider similarities with Bayesian interim monitoring using the predictive posterior probability (PPP), where the posterior probability may be analogous to the p-value and the PPP threshold analogous to our upstrapped proportion (12; 13; 14).

Based on our simulation studies, the upstrap has potential to serve as a nonparametric approach to implementing interim analyses in clinical trials. This is especially true for futility monitoring, with more work being needed for efficacy monitoring to better balance the type I error rate and power trade-off beyond the reduction in the expected sample size. In practice, the upstrap is flexible and could be generalized to a range of outcome types and study designs and is worth considering in future clinical trial designs for interim monitoring.

References

[1] DeMets, D. L., & Lan, K. K. (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine, 13*(13-14), 1341–1356. https://doi.org/10.1002/sim.4780131308

[2] K.K. Gordon Lan, David M. Reboussin & David L. DeMets (1994) Information and information fractions for design and sequential monitoring of clinical trials, Communications in Statistics - Theory and Methods, 23:2, 403-420, DOI: 10.1080/03610929408831263

[3] Jennison, C. and Turnbull, B. W. (1999). Group sequential methods with applications to clinical trials. CRC Press.

[4] O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*(3), 549–556.

[5] Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika, 64*(2), 191–199. https://doi.org/10.2307/2335684

[6] Pocock S. J. (1982). Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics, 38*(1), 153-62.

[7] Proschan M. A., Lan K. G., Wittes J. T. (2006). Statistical monitoring of clinical trials: a unified approach. Springer Science & Business Media..

[8] Kaizer, A. M., Wild, J., Lindsell, C. J., Rice, T. W., Self, W. H., Brown, S., Thompson, B. T., Hart, K. W., Smith, C., Pulia, M. S., Shapiro, N. I., & Ginde, A. A.

(2022). Trial of Early Antiviral Therapies during Non-hospitalized Outpatient Window (TREAT NOW) for COVID-19: a summary of the protocol and analysis plan for a decentralized randomized controlled trial. *Trials, 23*(1), 273. https://doi.org/10.1186/s13063-022-06213-z

[9] Kaizer, A. M., Shapiro, N. I., Wild, J., Brown, S. M., Cwik, B. J., Hart, K. W., Jones, A. E., Pulia, M. S., Self, W. H., Smith, C., Smith, S. A., Ng, P. C., Thompson, B. T., Rice, T. W., Lindsell, C. J., & Ginde, A. A. (2023). Lopinavir/ritonavir for treatment of non-hospitalized patients with COVID-19: a randomized clinical trial. *International Journal of Infectious Diseases, 128*, 223-229.

[10] Crainiceanu, C. M., & Crainiceanu, A. (2020). The upstrap. *Biostatistics (Oxford, England), 21*(2), e164–e166. https://doi.org/10.1093/biostatistics/kxy054

[11] Alsouqi, A., Deger, S. M., Sahinoz, M., Mambungu, C., Clagett, A. R., Bian, A., Guide, A., Stewart, T. G., Pike, M., Robinson-Cohen, C., Crescenzi, R., Madhur, M. S., Harrison, D. G., & Ikizler, T. A. (2022). Tissue Sodium in Patients With Early Stage Hypertension: A Randomized Controlled Trial. *Journal of the American Heart Association, 11*(8), e022723. https://doi.org/10.1161/JAHA.121.022723

[12] Zabor E. C., Kaizer A. M., Garrett-Mayer E., & Hobbs B. P. (2022) Optimal sequential predictive probability designs for early-phase oncology expansion cohorts. *JCO Precision Oncology*(6), e2100390. https://doi.org/10.1200/PO.21.00390

[13] Dmitrienko A. & Wang M. D. (2006) Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in medicine, 25*(13), 2178-95.

[14] Saville B. R., Connor J. T., Ayers G. D., & Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials, 11*(4), 485-93.

		Difference in Rejection Rate from Fixed Sample			*Ratio of ESS to Fixed Sample*		
Method	**N**	**FO**	**EO**	**FE**	**FO**	**EO**	**FE**
CU	40	-0.007	0.059	0.039	0.48	0.98	0.45
CU	160	-0.011	0.081	0.060	0.46	0.94	0.43
CU	600	-0.015	0.094	0.064	0.45	0.94	0.41
CU	2000	-0.015	0.077	0.055	0.45	0.94	0.41
AU	40	0.001	0.187	0.181	0.58	0.88	0.48
AU	160	-0.008	0.162	0.154	0.56	0.87	0.46
AU	600	-0.007	0.169	0.160	0.61	0.87	0.51
AU	2000	-0.005	0.180	0.175	0.62	0.86	0.51
GU	40	0.003	0.074	0.069	0.60	0.98	0.58
GU	160	-0.002	0.070	0.062	0.61	0.95	0.56
GU	600	-0.008	0.078	0.072	0.61	0.94	0.56
GU	2000	-0.010	0.071	0.061	0.59	0.95	0.55
OBF	40	0.004	0.006	0.004	0.68	1.00	0.68
OBF	160	0.001	0.003	0.001	0.69	1.00	0.69
OBF	600	0.000	-0.008	-0.003	0.71	1.00	0.70
OBF	2000	0.000	0.003	0.002	0.72	1.00	0.71
PO	40	-0.006	-0.010	-0.009	0.43	1.00	0.40
PO	160	-0.002	-0.004	-0.005	0.49	0.99	0.43
PO	600	-0.002	0.005	0.003	0.53	0.98	0.47
PO	2000	-0.001	-0.001	0.002	0.54	0.98	0.50

Table 1. *Null scenario* simulation results for the performance of the interim analysis calibration method relative to the fixed sample design. The rejection rate in the null scenario represents the type I error rate. FO is futility-only stopping, EO is efficacy-only stopping, FE is both futility and efficacy stopping. CU is the calibrated upstrap, AU is the arbitrary upstrap, GU is the group-sequential upstrap, OBF is the O'Brien-Fleming alpha-spending function, PO is the Pocok alpha-spending function.

		Difference in Rejection Rate from Fixed Sample			*Ratio of ESS to Fixed Sample*		
Method	**N**	**FO**	**EO**	**FE**	**FO**	**EO**	**FE**
CU	40	-0.118	0.086	-0.051	0.80	0.70	0.53
CU	160	-0.134	0.068	-0.074	0.80	0.63	0.46
CU	600	-0.144	0.050	-0.107	0.79	0.61	0.44
CU	2000	-0.185	0.055	-0.141	0.79	0.60	0.43
AU	40	-0.048	0.120	0.066	0.90	0.55	0.45
AU	160	-0.071	0.089	0.017	0.86	0.55	0.43
AU	600	-0.061	0.070	0.005	0.89	0.53	0.44
AU	2000	-0.068	0.065	-0.001	0.91	0.51	0.43
GU	40	-0.019	0.094	0.043	0.90	0.68	0.58
GU	160	-0.048	0.057	0.000	0.90	0.65	0.55
GU	600	-0.064	0.034	-0.019	0.90	0.61	0.52
GU	2000	-0.101	0.037	-0.049	0.89	0.61	0.52
OBF	40	0.017	0.008	-0.011	0.95	0.90	0.85
OBF	160	-0.012	0.012	-0.007	0.94	0.86	0.81
OBF	600	-0.018	-0.017	-0.021	0.95	0.84	0.79
OBF	2000	-0.023	-0.006	-0.022	0.96	0.83	0.79
PO	40	-0.122	-0.108	-0.187	0.75	0.85	0.60
PO	160	-0.092	-0.068	-0.163	0.83	0.79	0.56
PO	600	-0.094	-0.090	-0.151	0.84	0.76	0.57
PO	2000	-0.083	-0.084	-0.131	0.87	0.73	0.58

Table 2. *Alternative scenario* simulation results for the performance of the interim analysis calibration method relative to the fixed sample design. The rejection rate in the null scenario represents the power. FO is futility-only stopping, EO is efficacy-only stopping, FE is both futility and efficacy stopping. CU is the calibrated upstrap, AU is the arbitrary upstrap, GU is the group-sequential upstrap, OBF is the O'Brien-Fleming alpha-spending function, PO is the Pocok alpha-spending function.