Cooling to Help Injured Lungs (CHILL) Phase IIB Randomized Control Trial of Therapeutic Hypothermia in Patients with ARDS

Statistical Analysis Plan for the Primary Endpoint

Sponsor: University of Maryland through a grant from the Department of Defense

Protocol Title: Cooling to Help Injured Lungs (CHILL) Phase IIB Randomized Control Trial of Therapeutic Hypothermia in Patients with ARDS

Version/Date: 1.2/11-Aug-2023

Contents

1.0 List of Abbreviations and Definitions of Terms

APACHE II score = acute physiology score + age points + chronic health points.

ARDS: Acute Respiratory Distress Syndrome

CHILL: Cooling to Help Injured Lungs Clinical Trial

CONSORT: Consolidated Standards of Reporting Trials

CRRT: Continuous Renal Replacement Therapy

CSPCC: Cooperative Studies Program Coordinating Center

DCC: Data Coordinating Center

DSMB: Data Safety and Monitoring Board

eDC: Electronic Data Capture

ICU: Intensive Care Unit

LAR: Legally Authorized Representative

MOO: Manual of Operations

NMB: Neuromuscular Blockade

RCT: Randomized Clinical Trial

SAP: Statistical Analysis Plan

SOFA: Sequential Organ Failure Assessment

TH: Therapeutic Hypothermia

UTM: Usual Temperature Measurement

VFDs: Ventilator Free Days

WMW: Wilcoxon-Mann-Whitney

2.0 Introduction

This Statistical Analysis Plan (SAP) describes the analyses to be conducted for the primary endpoint of the Cooling to Help Injured Lungs Clinical Trial (CHILL).

2.1 Primary Aim

To test for the first time in a randomized clinical trial the hypothesis that early treatment with therapeutic hypothermia (TH) with neuromuscular blockade (NMB) to prevent shivering will be beneficial for patients with Acute Respiratory Distress Syndrome (ARDS).

2.2 Primary and Other Endpoints

2.2.1 Primary Endpoint

The primary endpoint will be a composite of rank-transformed 28-day VFDs for those who survive to day 28 and worst ranks assigned to those who die during the day 1 to day 28 period according to day of death with lower ranks assigned to those who die sooner. Twenty-eight-day VFDs is defined as the sum of all ventilator-free periods lasting at least 48 hours from day 1 to day 28. Ranks assigned using the above method have been given the term "worst rank scores".[7] Higher worst rank scores correspond to better outcome. Lung transplants will be treated as deaths in all analyses.

It is expected that few patients will be lost to follow-up or withdrawn from the study prior to day 28. For participants who are lost to follow-up or completely withdrawn from the study during the day 1 to day 28 period, vital status on day 28 as well as day of death (in case of death) will be determined, if necessary, from public records. The following procedure regarding the primary outcome will be followed for patients who withdraw completely from the study.

1. If the withdrawal occurs ≤ 54 hours from randomization and:
    a. the patient *did not* have an SAE resulting in death, the patient will be dropped from the primary analyses,
    b. the patient *did* have an SAE resulting in death, the patient will be included in the primary analysis (and assigned worst rank score as detailed above).
2. If the withdrawal occurs > 54 hours the patient will be dropped from the analysis.

We will perform two primary outcome sensitivity analyses. For each patient who completely withdrew from the study, we will repeat the primary outcome analysis after imputing the median worst rank score:

(1) from the treatment group to which the patient *was* assigned.
(2) from the treatment group to which the patient *was not* assigned.

If these sensitivity analyses return the same results as the primary analysis, the primary analysis will be supported. If the results differ markedly the primary analysis will be undermined.

2.2.2 Secondary Clinical Endpoints

Secondary Clinical Endpoints include: (a) 28-day ICU-FDs; (b) SOFA scores (trajectory over days 1-4 and 7); (c) 60- and 90-day functional status (ambulatory ability; use of supplemental oxygen; use of invasive or non-invasive ventilatory support (mechanical ventilation) for any part of the day except for CPAP, biPAP or variable (auto-titrating) PAP for obstructive apnea); (d) hospital, 60-day, and 90-day mortality. SOFA scores for days 0-3 will exclude the neurologic component when subjects in the TH arm will be receiving NMB.

2.3 Study Design

The study will be a randomized, open-label, parallel, two-group multicenter trial.

2.3.1 Study Population

The study population includes:

1. men and women
2. any race/ethnicity
3. 18 years, 0 days - 75 years, 364 days of age
4. endotracheal tube or tracheostomy in place and mechanically ventilated for ≤7 days
5. admitted to a participating ICU
6. radiologic evidence of bilateral pulmonary infiltrates
7. P/F ratio (partial pressure of oxygen ÷ fraction of inspired oxygen) ≤200 with PEEP (positive end-expiratory pressure) ≥8 cm H_2O. If ABG (arterial blood gas) values are not available, the P/F ratio may be inferred from SpO_2 (peripheral capillary oxygen saturation) values based on Table 3 from Brown et al. (see study protocol) as long as following conditions are met:
    a. SpO_2 values are 80-96%
    b. SpO_2 is measured ≥10 min after any change in fraction of inspired O_2 (F_IO_2)
    c. PEEP is ≥ 8 cm H_2O
    d. the pulse oximeter waveform tracing is adequate
    e. the qualifying inferred P/F ratio is confirmed 1-6 hrs after the initial determination.
8. For patients on APRV, the PEEP value will be calculated based on a formula (see protocol).
9. Patient is able to give consent, or a Legally Authorized Representative (LAR) is available to provide consent.

10. Criteria 6 and 7 must be met within 72h of enrollment and prior to randomization, not be fully explained by hydrostatic pulmonary edema. Criterion 6 must have occurred within 7 days of exposure to a condition associated with ARDS, including COVID-19 (note for COVID the 7-day window begins when patient is hospitalized for COVID-related illness).

*Patients may be enrolled and decision about randomization delayed if all criteria exclusive of P/F ratio ≤200 are met and then randomized if and when the P/F ratio ≤200 (as long as this occurs within the 72h ARDS window and 7-day mechanical ventilation window). If and when the P/F ratio criterion is achieved within the inclusion windows, the patient is eligible for randomization even if the P/F ratio subsequently exceeds 200 prior to randomization (e.g. while waiting for consent).

The study population excludes patients who have:

1. Missed moderate to severe ARDS window (>72hrs)
2. Missed NMB window (continuous NMB for at least 48 hrs prior to randomization: If is interrupted or discontinued before the patient exits the MV and ARDS windows, they are eligible for enrollment.)
3. Missed mechanical ventilation window (>7 days)
4. Refractory hypotension (requiring continuous administration of > 0.3 mcg/kg/min norepinephrine or equivalent dose of other vasopressors within 2 hrs of randomization). Note – this exclusion is time-dependent; if hypotension resolves before the patient exits the MV and ARDS windows, they are eligible for enrollment.
5. Spontaneous hypothermia for ≥ 6 hrs (core temperature <35.5°C while not receiving CRRT) on the day of randomization
6. Significant, active bleeding (>3U blood products and/or Surgical/Interventional Radiology intervention)
7. Platelets <10K/mm^3 (uncorrected) on day of randomization.
8. Active hematologic malignancy
9. Skin process that precludes cooling device
10. Moribund, not likely to survive 72h
11. Pre-morbid condition makes it unlikely that patient will survive 28 days
12. Do Not Resuscitate status (except for patients who are receiving full support except chest compressions)
13. Not likely to remain intubated for ≥48h

14. Physician of record unwilling to participate
15. Severe underlying lung disease
    a. On home O_2 ≥2LPM or ≥28%
    b. On home noninvasive ventilation (except for treating OSA)
    c. Prior lung transplantation
16. BMI >50 kg/m^2
17. Known NYHA class IV heart disease
18. Pregnant
19. Acute Coronary Syndrome past 30 days (MI, unstable angina)
20. Cardiac arrest within 30 days of enrollment
21. Burns over >20% of the body surface
22. Severe chronic liver disease (Child-Pugh score 12-15)
23. Previously enrolled in CHILL study
24. Simultaneous enrollment in another inpatient interventional trial initiated during the current hospitalization

2.3.2 Treatment Groups

Patients will be randomly assigned to either therapeutic hypothermia (TH) plus Neuromuscular Blockade (NMB) or usual temperature management (UTM).

2.3.3 Study Visits and Assessments

Assessments will be performed according to the schedule of events in the study protocol (Section 8.0). Section 13.0 of the MOO provides detailed lab and clinical values to be collected. Demographic data will be collected during screening. Baseline data will be collected prior to randomization and will include vital signs, clinical values, medical history, medications, NMB status, proning status. Blood samples will also be collected at baseline. Day 1 commences at randomization. On each of days 1-4 and 7, vital signs, patient status, proning status, NMB status, and other clinical information, as well as other select clinical and lab values will be collected. On each of these days a blood sample will be collected. On day 28, 28-day VFDs and ICU-free Days will be calculated and captured in the eDC system. On days 60 and 90, the vital status of the patient is collected and if alive, data regarding his/her location, need for ventilation or oxygen, and information on how ambulating. Additional data will be collected when important unscheduled events occur: adverse events, serious adverse events, commencement of unassisted breathing (UAB), ICU discharge, hospital discharge, and death.

2.3.4 Randomization and Blinding

This randomized clinical trial is unblinded. The nature of the study intervention precludes blinding.

Only patients who meet all the inclusion criteria, have none of the exclusion criteria, have a signed Informed Consent Form by themselves or their LAR, and have had the required pre-randomization BASELINE data and the BASELINE research blood sample secured (Forms F01, F06, F07, F12b, F17b, and F18) will be randomized.

The COVID-19 pandemic has resulted in a preponderance of ARDS patients infected with COVID-19. Although there are reasons to believe that the pathophysiology of organ injury, including ARDS, may be different in COVID-19 patients compared with other ARDS patients, we acknowledge that ARDS is a syndrome with diverse precipitating factors including COVID-19. Therefore, we will include patients with ARDS occurring in the setting of COVID-19 infection. Note that enrolling COVID-19 patients is not mandatory. COVID-19 positive and negative patients will be combined for randomization purposes.

A separate randomization schedule will be prepared for each site. Randomization will be performed within each site with study subjects to receive TH or Usual Temperature Management (UTM) using a 1:1 assignment ratio. Patients are assigned to TH or usual temperature management treatment in random order within blocks so that an equal number are assigned to each treatment when each block is completed. The order of block sizes (2,4, or 6) is random with the probability of each block size specified by DCC staff.

Assignment of the subject to TH or UTM will be made by the web based CSPCC randomization service (using Medidata Solutions Inc.'s randomization module).

2.3.5 Sample Size and Power

Using PASS v. 11[11] sample size software, we estimated the needed sample size by simulating our composite endpoint and using the Wilcoxon-Mann-Whitney (WMW) test of treatment arm difference. We seek to have power to detect a 4-day reduction in 28-day VFDs, which was the reduction in 28-day VFDs achieved in the PROSEVA trial[1]. To simulate the worst ranks distribution in the control arm (and CHILL arm under the null hypothesis) we assumed a 28-day mortality rate of 32%, approximately equal to the rate in the control group of the PROSEVA trial and in the historical control group in our prior pilot study (n=58). We estimated another 32% of the control arm would survive to day 28 but have zero 28-Day VFDs (consistent with our pilot data) and that the distribution of 28-Day VFDs in the remaining 36% of the control arm would be uniform over the interval 5 to 25 days (as in our pilot study control group). For the CHILL arm we reduced the mortality rate and number with 28-Day VFDs = 0 by nine percentage points and increased the number of 28-Day VFDs among the remainder to simulate an overall 4-day increase in 28-day VFDs as achieved in the PROSEVA trial. With the above parameter settings,

power set to 90% and alpha-level set to two-sided 0.05, the initial sample-size estimate for the WMW test applied to worst ranks equaled 258 (129 per group).

The sample size was then corrected to allow 10% one-way crossover assuming subjects randomized to hypothermia may not receive study treatment because of technical issues but that subjects randomized to standard treatment would not likely receive hypothermia treatment. The 10% is conservative (resulting in a larger total N) as the cross-over is typically ~3-4% in similar trials. This adjustment resulted in a sample size equal to $129/(1\text{-}0.1)^2 = 160$ per group. Hence, the final total N = 320. Because this estimate is very close to a prior estimated sample size of 324 used in planning the study, we are retaining the final total N = 324.

3.0 General Considerations for Data Analyses

3.1 Analysis Populations

3.1.1 Intention to Treat Population

The intention-to-treat population includes all patients randomized, analyzed by treatment assigned at randomization.

3.1.2 Per-Protocol Population

Patients in the per-protocol population will be analyzed according to the treatment received. The per-protocol population will include only randomized participants who are alive 54 hours post randomization. Patients in the per-protocol population must also have either completed 28 days of follow-up with assessed 28-Day VDS or have died within 28 days of randomization. The per-protocol UTM control population is defined as subjects randomized to the control arm whose core temperature is *not* ≤ 35.0°C for 24 hours or more during the first 54 hours post randomization. The per-protocol TH population is defined as subjects randomized to TH whose core temperature is ≤ 35.0°C for at least 24 hours within the first 54 hours post randomization.

The number of hours ≤ 35.0°C during the first 54 hours will be totaled summing the recorded 2-hour temperature blocks in the CRF. For blocks with both a high and low temperature, each of the two temperatures will be assumed to represent 1 hour (e. g. for a control patient a 2-hour block with the low temperature of 34.9°C and a high temperature of 35.5°C would count as one hour below 35.0°C). For a block with only a single temperature, that temperature will be assumed to represent 2 hours. For blocks that have no temperatures entered, the temperature of the blocks will be imputed from the mean temperature of the adjacent blocks. Temperature will not be imputed If the first or last block is missing.

The TH intervention procedure directs treating clinicians to cool patients down to a target temperature 34°-35.0°C and then maintain the patient's temperature in this range for the next 48 hours. The 54 hours in criterion (1) above come from the expectation that patients in the TH arm will be cooled to the target temperature range within 6 hours of randomization (see section 6.0 of the study protocol, 'post randomization period') and will receive therapeutic

hypothermia over the next 48 hours. For equitable comparison, the same 54-hour criterion is used for the UTM control arm.

The definition for cross-over aligns with the above definition of the per protocol population. Cross-over to the UTM arm among those randomized to TH is defined as subjects who *fail* to be cooled ≤ 35.0°C for at least 24 hours during the first 54 hours post randomization. Cross-over to the TH arm among those randomized to UTM is defined as subjects who have a core temperature ≤ 35.0°C for at least 24 hours during the first 54 hours post randomization.

3.2 Statistical Analysis Issues

3.2.1 Strata and Covariates

The randomization is stratified only by site, which will be accounted for in the primary efficacy analysis.

3.2.2 Multiple Comparisons

For the primary analysis, we consider a two-sided p-value of 0.044 to be significant. It is less than 0.05 to account for the interim analyses. For secondary and exploratory analyses, we consider $p<.05$ statistically significant and will not adjust for multiple tests to control the type I error rate. We consider $p<0.001$ as strong evidence.

3.2.3 Multi-Center Studies

There are a total of 14 sites participating in the study.

3.2.4 Study Baseline and Visits

Study baseline corresponds to measurements collected at the baseline visit, prior to randomization and initiation of treatment. Follow-up visits after hospital discharge are scheduled based on the date of randomization.

4.0 Interim Analysis and Data Safety and Monitoring Board

The CHILL Data and Safety Monitoring Board (DSMB) will review the accumulating data for early, convincing evidence of benefit or harm. We anticipate DSMB reports at approximately six-month intervals over the period of patient recruitment and follow-up. There will be three interim analyses followed by the final analysis after the conclusion of data collection. The three interim analyses will be performed after approximately 25% (n=81), 50% (n=162) and 75% (n=243) of subjects have either been followed-up for 28 days or have died prior to the 28th day. The primary statistical analysis method (stratified Wilcoxon-Mann-Whitney test comparing worst ranks) will be the same for the interim analyses as for the final analysis. For interim analyses, worst ranks for study withdrawal and loss-to-follow-up cases will be assigned in the same way as for the primary analysis as defined above in section 2.2.1 (Primary Endpoint). As the primary statistical analysis method requires each site to have randomized at least three

patients and at least one to each treatment group, all sites that haven't met this requirement will be combined into a single conglomerate site for analysis purposes. To inform the decision whether to stop the trial early due to the superiority of TH and maintain the overall Type I error rate (α) at 0.05 (two-sided) over repeated tests, we propose to use the Lan-DeMets[2] alpha-spending function approach with approximate O'Brien-Fleming-like[3] upper boundary. The spending function is: $\alpha(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t^*})$ where $t^* = [n_{CH}^{-1} + n_{UTM}^{-1}]^{-1}/[N_{CH}^{-1} + N_{UTM}^{-1}]^{-1} = [n_{CH}^{-1} + n_{UTM}^{-1}]^{-1}/81$ and n_{CH} and n_{UTM} equal to the interim sample sizes of the two arms and total $N = N_{CH} + N_{UTM} = 162 + 162 = 324$.[4] As calculated using PASS v.11, the z critical values (cv's) that much be equaled or surpassed by the stratified WMW z test statistic are provided in the table below under Upper Boundary for t* = .25, .50, and .75. To account for sequential testing, rejecting the null hypothesis favoring TH after completion of the study requires the Z test statistic to surpass 2.01406 in the primary efficacy analysis. To inform the decision to stop the trial early due to harm, we propose to use the approximate O'Brien-Fleming-like lower boundary as provided in the table below. Final boundaries will be adjusted as necessary if the interim analyses are not conducted at exactly 81, 162, and 243 cases, and/or if the final analysis is not at exactly 324 cases.

Table. Approximate Upper and Lower O'Brien-Fleming-Like Boundaries (PASS v. 11[10])

Look	t*	Lower Boundary	Upper Boundary	Nominal Alpha	Alpha Increment	Total Alpha
1	0.25	-4.33263	4.33263	0.000015	0.000015	0.000015
2	0.50	-2.96311	2.96311	0.003045	0.003036	0.003051
3	0.75	-2.35902	2.35902	0.018323	0.016248	0.019299
4	1.00	-2.01406	2.01406	0.044003	0.030701	0.050000

We also propose that the trial could be stopped early if there is evidence of futility based on the conditional probability of rejecting the null hypothesis given the "current trend" is less than 20% at the 2nd interim analysis or 3rd interim analysis. Conditional power will be calculated by normal approximation as suggested by Lan and Wittes[5] (1988) using the formula provided on page 1026 of Chen, DeMets, & Lan (2004).[6] Before making a recommendation to terminate the trial, either for a beneficial effect, harm, or futility, the DSMB will consider the interim analysis results along with other relevant factors. These factors will include enrollment and study conduct, the numbers and distributions of deaths, the frequencies and distributions of other secondary outcomes, and findings from other relevant studies

5.0 Subject Disposition

5.1 Disposition of Participants

A CONSORT diagram (see http://www.consort-statement.org/) will depict the disposition of all participants screened and randomized.

5.2 Extent of Exposure

To examine extent of exposure we will summarize the distribution of high and low temperature readings (or single temperature reading if only one measurement), measured within every 2-hour block, by treatment arm. Summaries will be both graphical (side-by-side boxplots) and numerical (select percentiles, for example, 90^{th}, 95^{th}, and 99^{th} percentiles for high temperature). Temperature distributions will be examined at baseline and over the 54 hours after randomization. Summaries will be examined stratified by the two-hour blocks and collapsed over time. The 54-hour time-period is defined based on the up to 6-hour time-period needed to reach target temperature (34°-35°C) in the TH arm plus the subsequent 48-hours of therapeutic hypothermia. In the TH arm, the percent of 2-hour blocks with high or single temperature less than or equal to 35.0°C over the 48-hours after target temperature is first achieved will also be calculated.

5.3 Protocol Violations

Protocol violations will be reported for the intention-to-treat population using counts and percentages of violations.

6.0 Baseline Data

6.1 Demographics and Baseline Characteristics

Baseline characteristics (pre-neuromuscular blockade) by treatment group for the intention-to-treat population will be compared using counts and percentages for categorical variables. Either the mean and standard deviation or median and interquartile range (25^{th}/75^{th} percentiles) will be reported for continuous variables. The following baseline characteristics will be reported: Age, gender, race, ethnicity, APACHE II, P/F ratio, proning status (at baseline), steroids, shock, SOFA, PEEP (cm H_2O), renal replacement therapy, COVID, and IL-6, bicarb, and protein C biomarkers.

7.0 Efficacy Analyses

7.1 Primary Efficacy Analysis

Primary and secondary analyses will be performed on the intention-to-treat (ITT) population. Per protocol analyses will also be performed as part of assessing the feasibility of a Phase III clinical trial with a longer-term mortality outcome. The primary analysis will be to test the treatment group difference on the primary endpoint, worst ranks, while accounting for site in the analysis. This will be accomplished using Wilcoxon-Mann-Whitney rank sum test extended to account for stratification by site (known as the van Elteren test[8]). Worst ranks for participants who are withdrawn from the study or lost to follow-up will be assigned according to section 2.2.1 (Primary Endpoint). As for the interim analyses, any sites that haven't met the minimum sample size requirement (three randomizations, at least one to each treatment

group) will be combined together into a single conglomerate site for analysis purposes. The alpha level of the test will be 0.05 (two-sided).

7.2 Per-Protocol Analysis

As a secondary analysis, the same analysis as performed in the primary efficacy analysis will be performed on the per protocol population. In addition, we will compare treatment groups on the primary endpoint in the per protocol population with a regression model with worst ranks as the dependent variable and treatment group and treatment center and several prognostic covariates as independent variables. Treatment center (14 sites) will be specified as a random effect. The prognostic covariates will include proning status, age, race, gender, and steroid use. However, before fitting the above model we will first examine subgroup differences by adding interaction terms between treatment group and the other independent variables. First, we will include an interaction term between treatment group and center to test whether there is significant variance in treatment effect among centers. This interaction will be random since center is random and so the global likelihood ratio test statistic for testing the interaction would be compared to a weighted mixture of two chi-square distributions.[9] If significant ($p \leq 0.05$), we will retain this interaction term in the above regression model and estimate and report treatment effects by center (coded). Second, we will add the interaction term between treatment group and each of the above baseline prognostic variables and if any are significant, we will estimate and report treatment effects by identified subgroups.

7.3 Subgroup Analyses

As a secondary analysis of the intent-to-treat analysis population, heterogeneity of treatment effect on the primary endpoint across levels of baseline patient characteristics will be assessed with a global test of interaction using a linear regression model. Independent terms in this model will include treatment group, baseline characteristic variable, and the interaction between treatment group and baseline characteristic variable. Baseline characteristic can be either categorical or continuous. A significant interaction (two-sided $p \leq 0.05$)will be followed by an estimate of treatment effect and 95% confidence interval for each level of categorical subgroups. *A priori* baseline characteristics to be tested include: proning status, shock, COVID, P/F ratio, time between meeting criteria for moderate to severe ARDS and randomization, age, and baseline biomarkers IL-6, bicarb, and protein C.

8.0 Interpreting Analysis Results

Interpretation of the primary outcome analysis must be cognizant of CHILL as a Phase IIb clinical trial. The primary outcome is intermediary to the ultimate outcome of longer-term mortality. Effects on longer term mortality could be observed fortuitously and will be interpreted as a secondary outcome.

If the null hypothesis is rejected, the investigators have the responsibility to assure and explain to the medical community that bias is not the reason for rejecting the null hypothesis. If the investigators fail to reject the null hypothesis, they have the responsibility to assure that they

have performed a sensitive test and explain to the medical community their basis for believing that bias is not the reason for failing to reject the null hypothesis.

Interaction with clinical site. Influence of clinical site on treatment effects will be inspected. If there are outlier clinical sites, explanations for the inhomogeneity will be sought (e.g., differences in fidelity to the treatment protocol) and any differences or inhomogeneities found will be presented with the primary analysis.

Effect size. A recommended effect size measure after performing a Wilcoxon-Mann-Whitney test statistics is known as the "probabilistic index" (q) and is defined as the estimated probability that an individual randomly selected from the study population will have a superior outcome if assigned to the experimental treatment, in our case therapeutic hypothermia. An advantage of this measure is it is mathematically linked with the WMW test statistic. A disadvantage of this effect size measure is that it is not on a commonly used measurement scale familiar to clinicians. When the outcome is on a simple quantitative scale, group medians are commonly reported after a WMW test and are more easily interpreted. However, this may not be sufficient for composite outcomes that combine mortality with a continuous outcome measure such as worst ranks. We will, therefore, report q as an overall effect size with a 95% confidence interval, and supplement it with more clinically interpretable summary statistics of the two constituent components underlying our worst ranks primary endpoint: (1) the 28-day mortality rate in each group and (2) median 28-day VFDs among the 28-day survivors in each group. Reporting q with a 95% confidence interval after performing the equivalent of a WMW test of worst ranks and the above two summary statistics of the constituent components was the recommendation for ARDS trials in a recent article in *Critical Care Medicine*.[11]

In addition, we will calculate the median worst rank score (including worst rank scores assigned to those who died before day 28) for each group and reverse transform these values to 28-Day VFDs according to the mapping of 28-Day VFDs to worst ranks. We refer to these as "median 28-Day VFDs based on worst ranks". Interquartile ranges will also be calculated in this way with the possible modification that the 25th percentile may correspond to death before day 28.

We will report the above effect-size measure and summary statistics in both the ITT and Per-Protocol populations. We will compare the WMW test result (p-value), effect size q, and constituent summary statistics between the two populations to further aid interpretation of the results.

References

1. Guerin, C. et al. Prone Positioning in Severe Acute Respiratory Distress Syndrome. *N Engl J Med* **368**; 23 (2013).
2. Lan, K. & DeMets, D. Discrete Sequential Boundaries for Clinical-Trials. *biometrika* **70**, 659-663 (1983).
3. O'Brien, P. C. & Fleming, T. R. A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556 (1979).

4. Halperin, M., Lan, K. K., Ware, J. H., Johnson, N. J. & DeMets, D. L. An aid to data monitoring in long-term clinical trials. *Control Clin Trials* **3**, 311-323 (1982).
5. Lan, K. K. & Wittes, J. The B-value: a tool for monitoring data. *Biometrics* **44**, 579-585 (1988).
6. Chen, Y.H.J, DeMets, D.L., Lan, K.K.G. Increasing the sample size when the unblinded interim result is promising. *Statist. Med.* **23**, 1023-1038 (2004).
7. Lachin, J.M. Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials* 20:408–422 (1999).
8. van Elteren, P. H. On the Combination of Independent Two Sample Tests of Wilcoxon. Bulletin of the Institute of International Statistics 37, 351–361 (1960).
9. Verbeke, G. & Molenberghs, G. Linear MIxed Models for Longitudinal Data. (Springer-Verlag, 2000).
10. PASS 2011 Power Analysis and Sample Size Software (2011). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.
11. Novack, V., Beitler, J.R., Yitshak-Sade, M., Thompson, B.T., Schoenfeld, D.A., Rubenfeld, G., Talmor, D., Brown, S.M. (2020). Alive and Ventilator Free: A hierarchical, composite outcome for clinical trials in the acute respiratory distress syndrome. *Crit Care Med* 48(2), 158-166 (2020).