Active Bathing to Eliminate Infection (ABATE Infection) Trial

NCT02063867

Statistical Analysis Plan

June 30, 2015

Pragmatic Clinical Trials Demonstration Project

Title: Decreasing Bioburden to Reduce Healthcare-Associated Infections and Readmissions

ABATE (Addressing **B**ioburden while **A**dmitted **T**o **E**liminate) **Infection Trial**

1.0 Randomization Plan

1.1 Randomization

Randomization will occur at approximately the 8th month of the baseline period. Each participating hospital will be notified of their placement at that time. This will be done because of the requisite 1-3 month period to submit and schedule intervention protocols for approval by relevant hospital committees which often meet monthly or quarterly. As per routine policy in all hospitals, no training or implementation activities may occur prior to obtaining requisite hospital committee approvals. This will allow approval to occur and appropriate training of staff to occur prior to the phase in period which will involve acquisition and introduction of intervention product.

While this study is one of the largest cluster-randomized trials of hospitals, simple randomization of 54 hospitals will not ensure balance of key variables by chance alone, and without blocking could even result in unequal numbers of hospitals in each arm. For example, with a naïve randomization, there would be a 9% chance of a 22-33 split, or worse. Thus, randomization will be stratified, with strata constructed to maximize the chance of balance for both baseline admission volume and the primary outcome, MRSA and VRE clinical cultures attributable to participating units. In addition, we will evaluate the possibility of constructing strata that balance additional variables, such as the mean comorbidity index (Romano score) of all patients in a hospital's participating units, the percent of patients bathing daily, and the baseline use of chlorhexidine and mupirocin.

Achieving balance on key features of the randomization units (in this case, hospitals) is a critical task in cluster-randomized trials, but little literature on it exists. Unlike individually-randomized trials, information about the clusters is often known in advance, but the number of clusters to be randomized can be relatively small. The existence of a priori data can mitigate the small numbers and help to obtain adequate balance through stratification. One attractive approach is to establish tuplets—matched sets (pairs, for a two-arm trial) – in which one member of each tuplet is assigned to each arm. Schemes for constructing tuplets need not be guided by theory. A formal approach would be to calculate the Mahalanobis distance between hospitals across all key variables and choose the set of tuplets with the minimum average distance. In this approach, we could standardize the variables, and then multiply by values calibrated to reflect any difference in the importance of balancing them. Other approaches are more ad hoc, such as prioritizing broad classes of balance on a key variable and making pairs within these strata based on lower-priority variables. However, there is no "best" method of tuplet construction, only sets that come closer to meeting the varied needs of each trial.

We will enhance methods to inform the choice of tuplet-construction scheme which we developed in the REDUCE MRSA trial,[1 2] and share them (see software sharing plan). One example method is to establish the pairs under several plausible tuplet-construction schemes, and use graphical methods to compare all possible realizations for balance between the arms under each scheme. For example, if two variables must be balanced, we could tentatively divide the sample into two groups under a tuplet construction scheme and then generate a scatterplot showing the between-arm absolute value of the mean difference for one variable on the x-axis and the second on the y-axis for each possible result of the randomization. We would then divide the groups again under the same scheme, and find another point on the scatterplot. Repeating many times would show the typical and distribution of balance under a scheme. Comparing the resulting scatterplots from each tuplet-construction scheme can reveal the relative risks of imbalance and benefits for balance accruing to each randomization scheme, in a practical sense. One tuplet construction method may result in generally close balance on one key characteristic and very variable balance on the other, while a competing scheme has good median balance on both characteristics, but where each has a long tail implying a few bad-luck assignments with poor balance.

We hope to consider balance on more than two factors, and for assessing the impact on balance in this case, we will use a parallel coordinates plot, a multivariate plot method. A simulated example is shown to the right in **Figure 1**. There we show a potential result of a single tuplet construction method. The variables shown are volume, baseline rate of an outcome, the baseline rate of chlorhexidine use, and baseline rate of bathing.

Each blue, red, green, or black line shows the mean difference between arms for all four variables for one potential realized randomization. The results show that a few randomizations, in blue, are relatively imbalanced on volume and outcome but balanced on chlorhexidine use and bathing, while a few others, in black, have the reverse pattern. The green and red realizations are approximately equally balanced across these variables. If we considered it more important to balance on volume and outcome, this would probably not be an ideal scheme.

As a final note, the statistical core advised us to consider the relative costs and benefits of strata of four, rather than tuplets, which are strata of two. There are sound statistical reasons to expect power to be slightly better with strata of four, although there is some debate on this point.[3] [4] However, the balance between the arms may be worse. The balance is of central importance, since balance ensures that the observed effect is not confounded—confounding requires that the confounder be out of balance between the arms. We will examine whether the gain in power is strong enough, and the loss of balance slight enough, to pursue the strata of four in place of tuplets.

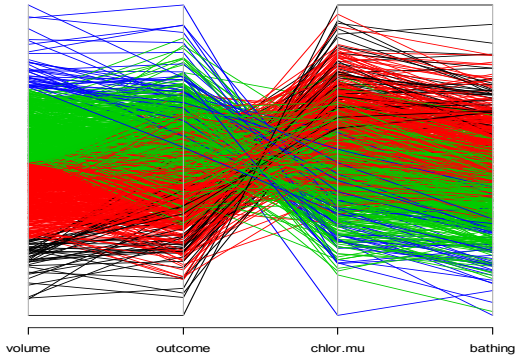

Figure 1. Parallel Coordinates Plot Showing Simulated Balance Across Multiple Variables

2.0 Statistical Analysis Plan

2.1 Finalized Outcomes

Study outcomes were finalized following deliberation of the Steering Committee during the UH2 planning year. Deliberations included response to recent published literature, including the REDUCE MRSA Trial which was conducted by our investigative team.1 2 The primary outcome will be the presence of at least one clinical culture with gram-positive multi-drug resistant bacteria (MRSA and VRE) attributable to a participating unit. This outcome was solidified following recent clinical trial evidence of the success of chlorhexidine (with and without mupirocin) in reducing these pathogens in ICUs.1 2 *An priori* secondary outcome intended for the primary manuscript is all-cause bloodstream infection attributable to a participating unit. Additional *a priori* secondary study outcomes intended for secondary manuscripts are provided in **Table 1.**

Table 1. Study Outcomes

Primary Outcome
MRSA and VRE clinical cultures [a]
Secondary Outcomes (Primary Manuscript)
All-cause bloodstream infection [a b]
Secondary Outcomes (Secondary Manuscripts)
Gram-negative multi-drug resistant organisms [a]
Urinary tract infections [a]
C difficile clinical tests [a]
Blood culture contamination
30-day infectious readmissions
Emergence of resistance to chlorhexidine or mupirocin [a]
Cost effectiveness

[a] Attributable to participating units. Defined as occurring >2 days into a participating unit stay through 2 days following unit discharge

[b] Includes subsets of GP and GN MDROs as well as key pathogens such as *S. aureus*

These outcomes are designed to maximize the evaluation of the impact of decolonization in non-critical care settings. They will address major concerns in healthcare related to reduction of antibiotic-resistant pathogens, and impact on a range of hospital-associated infections. They will also assess the likelihood that bacterial strains will develop resistance to chlorhexidine and mupirocin following broad use among inpatients.

2.2 Statistical Analysis

All outcomes will be assessed similarly. Here, we use the example of the primary outcome: clinical cultures with MRSA or VRE (first per patient). MRSA and VRE clinical cultures will be attributed to participating units if the collection date occurred >2 days after admission to that unit through two days after discharge from that unit. This attribution is consistent with CDC guidance for surveillance of nosocomial infections.[5]

Main trial results will be based upon as-randomized, unadjusted analyses using proportional hazards models to account for patients' variable tenure in the unit. This is necessary for the usual reasons: dichotomizing patients into those with vs. without infections would require us to define a fixed time-frame (within first x days of eligibility, ignoring some infections and omitting patients with shorter stays) or to ignore exposed time (counting infection during unit stay regardless of different length of stay). In addition, power is greater for proportional hazards models than for logistic regression models.[6 7 8 9]

Clustering within hospital will be accounted for using shared frailties. The frailties are added model terms that allow unique hazards ratios for each hospital, and are necessary to account for clustered randomization.[10] Model terms will include individual-level data on arm, hospital, outcome events, trial period (baseline vs. intervention) and an interaction term between trial period and arm. The assessment of trial success will be determined by the significance of the interaction term, which assesses whether the difference in hazard between the baseline and intervention period differs significantly between the two arms. We can write a simple version of the model symbolically as:

$$\lambda_{ij}(t) = \lambda_0(t)e^{\beta_1 Arm_{ij}+\beta_2 Period_{ij}+\beta_3 Arm_{ij}*Period_{ij}+\gamma_i}$$

where i is a hospital, j is a person within the hospital, Arm and $Period$ are indicator variables and are = 0 for patients in a hospital in the control arm or baseline period and 1 if in the intervention arm or period. The overall hazard rate over time for person ij is defined as $\lambda_{ij}(t)$, a function of the baseline hazard $\lambda_0(t)$, which is similar to the intercept in a linear model, times the proportionality for that subject, which is defined by the covariates Arm_{ij} and $Period_{ij}$ and the associated parameters, as well as the frailty, γ_i. The frailties are closely analogous to the random effect in generalized linear mixed models, and account for the clustering (similarity of hazard) within a given hospital. The ultimate effect of the intervention is assessed through β_3: as parameterized, if it is negative and has p-value < .05 (or 95% CI excluding 0) then the intervention reduces the risk of infection. We plan to assess the need for different frailties by hospital by period (instead of just by hospital), as well as additional clustering by unit within hospital. In addition, we will investigate the need to adjust for the stratified randomization scheme described in **Section 1.1.**

Subsequent analyses will include as-treated and covariate-adjusted models. Adjusted models will account for individual characteristics such as age, gender, comorbidities based upon ICD-9 codes, and receipt of intervention products. We will also account for unit type (step down, medical, surgical, etc.) and baseline bathing frequency if this is not balanced after randomization. All analyses will be performed using current versions of SAS (9.4, as of writing, SAS Institute, Cary NC) and/or R (3.0.2, as of writing).[11]

In addressing considerations of interim analyses to determine whether early stopping might be possible, the decision has been made not to pursue an interim analysis for several reasons. First, this trial meets the requirements of a minimal risk study. The study of topical bathing/decolonization therapy necessitates neither interim analyses nor stopping rules since reasonably anticipated adverse events are considered minor. Second, the collection time plus the lag in obtaining data and the relatively sparse power suggest that it is highly unlikely that an early look would result in a stoppage of the trial for either futility or success. Third, the addition of an interim analysis would affect power estimates in such a way to create an elongation of the trial beyond the time period that is acceptable to our health system partner. As described in detail above, participation in the trial requires a continuous assertion that other hospital interventions and campaigns that may conflict with the trial will not be pursued. This restriction to the usual tendency of hospitals to pursue multiple simultaneous interventions for prolonged periods of time was a critical consideration in designing the length and size of this trial. With regard to assessment of adverse events, we have built in a reporting system for both mild and severe side effects as described in our Data Safety Monitoring Plan and in our Decolonization Educational Materials.

2.3 Power

In many settings, an analytic approach to power is possible: given the assumptions of the model (e.g., logistic regression) are met, a relatively simple closed-form solution exists. However, generating the expected values to plug in may be difficult. In addition, some settings are complex enough that closed-form solutions may be difficult to generate. Many cluster-randomized designs fall into this class. In cluster-randomized

problems, it is also difficult to obtain reliable estimates of the additional parameters that are required, most notably the between-cluster variance or, equivalently, the intra-class correlation coefficient. Further additional complications are introduced for time-to-event outcomes such as those needed in the ABATE Infection Trial.

In a previous trial, we used the logistic regression analogue to proportional hazards regression models and simulation to estimate power.[7] Now, however, we propose an interesting and, to our knowledge, novel approach to power calculation, which we dub "bootstrap power calculation." This method is described below and in an article published since the trial was planned.[12] Briefly, the bootstrap is a powerful technique that uses the observed sample to approximate the underlying population, rather than a convenient analytic distribution.[13] The bootstrap power approach relies on the fact that we possess a large quantity of baseline data already. Loosely put, we (1) bootstrap a sample of observations from our observed baseline data to serve as the baseline sample in the power calculations. Then we (2) bootstrap another sample from the baseline data to serve as the intervention period data. Next, (3) we implement the randomization scheme in the bootstrapped sample. Then (4), for a randomly selected subset of the outcomes (e.g., bloodstream infections) observed in the bootstrapped intervention period sample in the hospitals randomized to intervention, we artificially change the outcome from infection to no infection. This represents the effect of the intervention, which we control by changing the size of the subset selected for this change. Simultaneously (5), for this subset we change the date of the event from the date of infection to the date of discharge, transfer to a non-eligible unit, or death—i.e., the date of censoring, had no infection been observed. Finally (6), we fit the planned frailty model described in **Section 2.2** above and record whether the null hypothesis of no association was rejected or not. This process is repeated many times, and the proportion of rejections is an estimate of the power under the given effect size. A confidence limit on this estimated power can be generated. In the table below, we show the power for removing the outcome from 0, 10, 20, and 30% of the subjects in the intervention arm in the intervention period. Removing 0% of the outcomes is a test of the technique: since we are not reducing the infection rate, the null is true, and rejections should occur only about 5% of the time.

In the initial planning of the trial, we used the available 4-month baseline sample of patients from our participating trial hospitals and used three bootstrap samples to represent each 12-month baseline and 4.5 bootstrap samples to represent the 18-month intervention period. This resulted in the power shown in **Table 2**.

Table 2: Power and Exact 95% CI for Primary and Select Secondary Outcomes*

Intervention Effect	Primary Outcome MRSA, VRE Clinical Cultures	Gram Negative MDRO** Clinical Cultures	All Pathogen Bacteremia
*0%	5.6% (4.3-7.2%)	4.1% (3.0 – 5.5%)	5.2% (3.9 – 6.8%)
10%	33% (28-37%)	15% (12 – 18%)	23% (19 - 27%)
20%	92% (90-94%)	44% (40 – 49%)	68% (63 - 72%)
30%	100% (99-100%)	82% (79 – 86%)	98% (96 – 99%)

*Based on 500 bootstrap samples for each effect size, except for the 0% estimate, for which we did 1000 simulations to increase confidence that the alpha level is maintained when the null is true.

** Gram-negative multi-drug resistant organism clinical cultures

The above power estimates (**Table 2**) show that the technique has the desirable characteristic of rejecting the null only 5% of the time when it is true. We also see that power for MRSA or VRE clinical culture is ample, even if we prevent only 20% of the infections. For both selected secondary outcomes, the power is less, but still quite acceptable if the intervention prevents 30% of the infections. The primary strengths of the bootstrap power approach are that it allows us to avoid using literature estimates for the parameters when such estimates may not apply to the trial population, and it also avoids unrealistic assumptions about regularity (equal cluster sizes) or distribution (logistic instead of frailty models). The main weakness in this case is that correlation within hospital is generated to be the same in the baseline and intervention periods. In addition, we have not received all desired randomization stratification data by the time of this submission. Thus, stratification in the bootstrap process is limited to hospital size and the baseline rate of the outcomes. We believe that these are relatively benign issues: the correlation structure is unlikely to change importantly from period to period during the actual study, and the stratification is mainly to promote balance. It may affect the power, but mainly by reducing the variability in the outcome.

After the baseline period, we were able to re-estimate the power, using the data collected in the full baseline period. We used the bootstrap process outlined above, except that we used the observed 12-month

data for the baseline. We assessed the power for the original proposal of an 18-month intervention and for a slightly extended 21-month intervention period, oversampling the observed 12-month baseline data as needed in each case (**see Table 3**). This updated power assessment was considered to be more accurate in that it used approximately three times as much real data—12 months vs. four months—compared to the original estimate. Furthermore, additional cleaning steps were applied to the 12 months of baseline data. In these new assessments, we focused on an intervention effect of 20%. The power for the primary outcome was 99.9% (95% CI: 99.4-99.99%) with 18 or 21 months of follow-up. For all pathogen bacteremia, the power was 85% (95% CI: 83-87%) with 18 months follow-up and 89% (87-90%) with 21 months follow-up. The steering committee decided to use 21 months of follow up to ensure greater power with less common secondary outcomes and account for lessened effects at sites that may drop out of the trial.

Table 3: Revised Power and Exact 95% CI for Primary and Select Secondary Outcomes*

Analysis	Effect	Primary Outcome MRSA, VRE Clinical Cultures	All Pathogen Bacteremia
As-randomized, 18-mo	20%	99.9% (99.4% – 99.99%)	85% (83% – 87%)
As-randomized, 21-mo	20%	99.9% (99.4% – 99.99%)	89% (87% – 90%)

Bibliography and References Cited

[1] Randomized Evaluation of Decolonization versus Universal Clearance to Eliminate MRSA (REDUCE MRSA) Cluster Randomized Clinical Trial http://clinicaltrials.gov/ct2/show/NCT00980980. Funded by Agency for Healthcare Research and Quality (Task Order PI: **Huang**; DEcIDE Network PI: **Platt**). Last accessed on April 19, 2012.

[2] Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery T, Lankiewicz J, Gombosev A, Terpstra L, Hartford F, Hayden M, Jernigan JA, Weinstein R, Fraser VJ, Haffenreffer K, Cui E, Kaganov RE, Lolans K, Perlin J, Platt R. Randomized evaluation of decolonization vs. universal clearance to eliminate methicillin-resistant Staphylococcus aureus in ICUs (REDUCE MRSA Trial, Abstract 36049). IDWeek (1st Annual Joint Meeting of IDSA, SHEA, HIVMA, and PIDS), October 17-21, 2012 (San Diego, CA)

[3] Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. Statistics in Medicine 1997;16(15):1753-64.

[4] Imai K, King G, Nall C. The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. Statistical Science 2009;24:29–53.

[5] Centers for Disease Control and Prevention. National Healthcare Safety Network. Tracking Infections in Acute Care Hospitals/Facilities. http://www.cdc.gov/nhsn/acute-care-hospital/index.html. Last accessed on September 28, 2013.

[6] Cuzick J. The efficiency of the proportions test and the logrank test for censored survival data. Biometrics 1982;38:1033–1039.

[7] Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Stat Med 1989;8:1515–1521.

[8] Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. J Chronic Dis 1983;36:715–723.

[9] van der Net JB, Janssens AC, Eijkemans MJ, Kastelein JJ, Sijbrands EJ, Steyerberg EW. Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. Eur J Hum Genet. 2008;16(9):1111-6. Epub 2008 Apr 2.

[10] Hayes RH, Moulton LH. Cluster Randomized Trials. CRC Press, New York 2009, p 207.

[11] R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. Last accessed September 28, 2013.

[12] Kleinman K, Huang SS. Calculating Power by Bootstrap, with an Application to Cluster-Randomized Trials. EGEMS (Wash DC). 2017 Feb 9;4(1):1202. doi: 10.13063/2327-9214.1202. eCollection 2016.

[13] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.1993. ISBN 0-412-04231-2