Project CLEAR – Statistical Analysis Plan
Lead Statistician: Daniel Gillen, PhD

1.0 Statistical Analyses

1.1 Data Processing and Statistical Analysis

All final statistical analyses will be completed in using R statistical software and results will be validated using SAS version 9.2 (Cary, NC).

1.2 General Analytic Considerations

The study population description will include demographic and socioeconomic characteristics, comorbidities, MRSA risk behaviors, recent surgical procedures, wounds or medical devices, personal hygiene activities, and household and employment characteristics at the time of study entry (hospital discharge). All descriptions will be stratified by study arm. Further exploratory descriptive statistics will also be presented by randomization (nursing home status and Hispanic race/ethnicity). Participant characteristics will be described by mean and standard deviation, by median and interquartile range, or by frequency and percentage for normally distributed variables, non-normally distributed variables, and categorical variables, respectively. We will use *t* tests, Mann-Whitney U tests, or χ^2 tests to compare these characteristics between the two study arms.

1.3 Analysis Populations

Safety analyses will be performed on the ITT Population, defined as all subjects who were randomized into the study. Efficacy analyses will be conducted on the Intention-To-Treat (ITT) Analysis Population, defined as all subjects who were deemed eligible and randomized in the study. Data summaries will be based on the intervention that was randomly assigned.

1.4 Handling of Missing Data

In primary analyses, missing values will not be imputed. All censored values will be considered in the primary analysis regardless of the reason for censoring.

1.5 Subgroup Analyses

Descriptive analyses of safety and efficacy data will be presented by Hispanic race/ethnicity and nursing home status. The primary efficacy endpoint, as well as all secondary efficacy endpoints and safety outcomes will be presented overall and by subgroup.

1.6 Accounting for Multiple Comparisons

The primary analysis comparing the relative risk of MRSA infection between the SOCE and SOCE+D arms will be conducted to maintain an overall 0.05 (two-sided) significance level. The primary analysis will be conducted in the full ITT Analysis Population. Subgroup analyses in Hispanic patients and nursing home residents are considered secondary and hence no adjustment for multiple comparisons will be accounted for in

the primary analysis.

1.7 Data Transformations

No transformations of data are planned.

1.8 Efficacy Data

1.8.1 Analysis Population

Efficacy analyses will be conducted on the Intention-To-Treat (ITT) Analysis Population, defined as all subjects who were deemed eligible and randomized in the study.

1.8.2 Analysis of the Primary Outcome

For the primary analyses, we will evaluate **time from discharge to first MRSA infection.** The primary endpoint will be summarized using Kaplan-Meier estimates of survival distributions and the 3 month, 6 month, and one year incidence of MRSA infection. Inference for the primary endpoint will be assessed using a Cox proportional hazard model. The proportional hazard assumption will be assessed graphically by plotting log(-log(estimated survival distribution function)) against log(survival time). The resulting graphs should have approximate parallel lines when the assumption holds. In addition, we will also assess potential treatment by time interactions. If the proportional hazards assumption is reasonably met, then the hazard ratio estimated via the Cox model will be used as summary of treatment effect. If the proportional hazards assumption is violated, then the inference remains statistically valid for testing equality in survival distributions, but treatment effect will be summarized by the proportion of observed infections at relevant time points.

1.8.3 Analysis of Secondary Outcomes

Additional secondary analyses will involve the assessment of MRSA infection and hospitalization rates. Given that **time to hospitalization associated with MRSA infection** reflects a very important high cost and high severity metric this outcome will also be assessed using the above Cox analysis. In addition, we will consider the impact of intervention on multiple re-infections/re-admissions. In these analyses, participants will be allowed to contribute multiple discrete MRSA infections. As above, we will provide descriptive data including the median time to infection, types of infection, and whether infections resulted in hospitalization, stratified by study arm. A multivariate Cox proportional hazards model will be used to estimate the relative hazard for re-infection and/or hospital readmission [LIN, 2000]. Separate baseline hazard functions will be estimated conditional upon the number of previous infections/admissions and multiplicative interactions will be used to test whether the effect of treatment differs by the previous number of infections/hospitalizations. Robust standard errors will be used for all inference in order to account for correlation within individuals with multiple infections/readmissions [LIN, 1989]. Secondary analyses will compare the rates of MRSA infection for a) both inpatient and outpatient infection combined, and b) only inpatients visits due to or complicated by MRSA infection

In addition, we will evaluate whether either arm differentially engendered antibiotic resistance. This will be done by comparing the proportion of patients in each arm with MRSA isolates obtained after at least six months of follow-up who has strains that were

non-susceptible to a) mupirocin, or b) chlorhexidine. Comparisons will be done using chi-square tests. Total deaths and deaths due to MRSA will also be described between the two study arms and compared using chi square tests.

All analyses will be repeated using as treated data where participants in the serial decolonization arm will be considered adherent to the study arm if the 5-day decolonization regimen is completed at least 50% of the time during the follow up period without more than two consecutive missed decolonization regimens. Participants in the education arm who receive any decolonization regimen will be excluded from this analysis.

Finally, we will construct data-driven classification models to identify groups at increased risk for short-term infection (within three months of discharge) and long-term infection (within one year of discharge) infection, as well as. Outcomes will be the MRSA the occurrence of infection within first 3 months following discharge and within the first 12 months following discharge. Two methods will be used to build classification model and assess covariate importance in predicting short and long term risk of MRSA infection. The models will then be compared based upon their out of sample misclassification rates. First, we will you logistic regression to model the log-odds of short- and long-term infection. To avoid model over-fitting and reduce out-of-sample classification error, parsimonious models will be chosen using 10-fold cross-validation, in which a proposed model will be fit using 9/10 of the available data and used to predict the remaining 1/10 of the sample. The resulting logistic regression will be summarized in terms of the risk score associated with coefficient values derived from the estimated model parameters. In addition, classification and regression trees (CART) will also be used to build a flexible predictive model, accounting for potential effect modification across covariates. Again, 10-fold cross-validation will be used for pruning CART trees to avoid overfitting. Using the final CART model as a guide for covariate selection, logistic regression will be used to model the odds of 3-year cognitive impairment and risk scores for covariate values will be derived from the fitted regression coefficients. Both the logistic regression and CART prediction will summarized via receiving operating characteristic curves and individual covariate contributions to the models will be reported.

1.9 Sample Size Justification

Initial power analyses considered exponential survival with a baseline one year infection rate of 24%. A blinded reassessment of power (pooling data over both treatment arms) was performed one year into the trial to more accurately assess the precision of the study based upon the overall event rates occurring in the trial. In the reassessment, power for the study was based upon the following design assumptions:

1. Level 0.05 (two-sided) for the primary outcome of MRSA re-infection
2. 24% baseline MRSA infection rate at 1 year
3. 75% of events infections will occur within 6 months of discharge
4. 5% of treated population will have high level resistance

Further we considered the lost-to-follow-up rate of 30% distributed uniformly from 0 to 2 months (was assumed to be 10% uniformly distributed over 12 months) and that there existed a decrease in initial treatment effect that attenuates to the null over the final 6 months of follow-up. Based upon these assumption, simulated power as a function of the hypothesized treatment effect (HR) for MRSA-specific infection (24% baseline event rate at 1 year) over the first 6 months of follow-up and an assumed relative reduction in the initial effect for the final 6 months of follow-up are provided in Table 2. A similar table

considering all infections (assuming a baseline infection rate of 35%) is provided in Table 3.

Table 2. Simulated power as a function of the hypothesized treatment effect (HR) for MRSA-specific infection (24% baseline event rate at 1 year) over the first 6 months of follow-up and an assumed relative reduction in the initial effect for the final 6 months of follow-up.

Percent Decrease in Effect for 6-12 Months	True Hazard Ratio For First 6 Months			
	0.60	0.65	0.70	0.75
N=1800				
50%	0.880	0.779	0.619	0.432
60%	0.849	0.750	0.596	0.408
70%	0.831	0.718	0.576	0.388
80%	0.808	0.688	0.545	0.371
90%	0.769	0.657	0.521	0.349
100%	0.733	0.628	0.497	0.317
N=1900				
50%	0.891	0.788	0.650	0.491
60%	0.861	0.760	0.625	0.465
70%	0.838	0.733	0.595	0.435
80%	0.814	0.703	0.562	0.419
90%	0.796	0.677	0.524	0.391
100%	0.777	0.647	0.493	0.368
N=2000				
50%	0.913	0.813	0.659	0.477
60%	0.885	0.788	0.633	0.456
70%	0.860	0.762	0.603	0.433
80%	0.834	0.732	0.585	0.411
90%	0.804	0.701	0.558	0.386
100%	0.778	0.666	0.533	0.360
N=2100				
50%	0.923	0.848	0.703	0.512
60%	0.899	0.819	0.674	0.493
70%	0.879	0.787	0.654	0.471
80%	0.867	0.758	0.614	0.449
90%	0.839	0.727	0.571	0.422
100%	0.811	0.685	0.546	0.389
N=2200				
50%	0.929	0.859	0.702	0.493
60%	0.914	0.833	0.688	0.468
70%	0.899	0.803	0.662	0.442
80%	0.875	0.780	0.634	0.415
90%	0.846	0.751	0.608	0.389
100%	0.828	0.712	0.576	0.371
N=2300				
50%	0.946	0.861	0.718	0.557
60%	0.933	0.840	0.686	0.532
70%	0.922	0.819	0.661	0.515
80%	0.899	0.786	0.625	0.491
90%	0.880	0.755	0.591	0.461
100%	0.862	0.726	0.565	0.434

Table 3. Simulated power as a function of the hypothesized treatment effect (HR) for <u>any infection (35% baseline event rate at 1 year)</u> over the first 6 months of follow-up and an assumed relative reduction in the initial effect for the final 6 months of follow-up.

Percent Decrease in Effect for 6-12 Months	True Hazard Ratio For First 6 Months			
	0.60	0.65	0.70	0.75
N=1800				
50%	0.979	0.920	0.795	0.594
60%	0.973	0.903	0.777	0.574
70%	0.965	0.889	0.749	0.545
80%	0.953	0.868	0.715	0.516
90%	0.938	0.841	0.690	0.474
100%	0.913	0.815	0.653	0.443
N=1900				
50%	0.976	0.932	0.804	0.635
60%	0.963	0.914	0.773	0.608
70%	0.958	0.888	0.740	0.575
80%	0.944	0.870	0.710	0.546
90%	0.936	0.850	0.675	0.527
100%	0.927	0.821	0.648	0.500
N=2000				
50%	0.984	0.946	0.828	0.653
60%	0.980	0.934	0.796	0.629
70%	0.974	0.913	0.767	0.605
80%	0.961	0.893	0.749	0.581
90%	0.946	0.869	0.711	0.553
100%	0.924	0.836	0.681	0.524
N=2100				
50%	0.990	0.938	0.838	0.662
60%	0.987	0.925	0.814	0.638
70%	0.982	0.905	0.788	0.621
80%	0.976	0.883	0.768	0.597
90%	0.966	0.865	0.752	0.565
100%	0.953	0.826	0.720	0.533
N=2200				
50%	0.986	0.971	0.831	0.689
60%	0.978	0.962	0.817	0.672
70%	0.973	0.948	0.793	0.647
80%	0.963	0.936	0.768	0.624
90%	0.948	0.917	0.745	0.586
100%	0.937	0.900	0.714	0.557
N=2300				
50%	0.990	0.964	0.892	0.716
60%	0.986	0.954	0.872	0.690
70%	0.984	0.940	0.838	0.658
80%	0.978	0.924	0.814	0.630
90%	0.970	0.901	0.781	0.602
100%	0.962	0.884	0.748	0.574

References

Lin D, Wei L, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of Royal Statistics. 2000. 62(4):711-30.

Lin D and Wei L. Robust inference for the Cox proportional hazards model. Journal of the American Statistical Association. 1989. 84:1074-8.