12. STATISTICAL CONSIDERATIONS

12.1. Study Hypotheses

Primary hypothesis. Null hypothesis is that the proportion of patients with a successful outcome of WHO OM severity grades 0-2 at the post study evaluation will not differ between the two randomized intervention groups. The estimated proportion of successes from our pilot trial is 0.30 with SOC. The alternative hypothesis is that the proportion of patients with a successful outcome differs between the two groups from the SOC proportion of 0.30 by +/-0.265 or more.

Secondary objectives will test whether or not there are differences in changes in cytokines levels over time by treatment group, and in other measurements, including time to onset and duration of severe OM (greater than 2), duration and time to onset of severe OM, salivary hypofunction, average mouth and throat soreness (MTS), and quality of life and function [EORTC QLQ 30 and EORTC QLQ HN35].

12.2. Sample Size Considerations

For the primary clinical study outcome of OM severity (WHO OTS scale) at end of RT/chemoRT, for simplicity, and for conservative sample size estimation, we consider subjects with grades 0, 1, 2 OM as having a successful outcome, and with OM grade greater than 2 as a failure. Subjects who never develop or scored 1 on the WHO mucositis scale will be considered as a success if they complete the end of treatment visit. The overall study will randomize 120 subjects in four strata. Sixty-five percent of eligible participants are expected to be treated with RT alone, the remaining 35% are expected to be treated with chemoRT. Approximately 45% of participants are expected to receive Proton therapy, 55% IMRT. The expected overall success rate on RT or chemoRT is estimated as 30% with SOC. With 1:1 randomization into the 2 treatment arms, with 80% power and 2-sided alpha of 0.05, we can detect a difference in this success rate of |26.5 %| based on Fisher's exact test for proportions (exact alpha = 0.03; 60 subjects per arm across the four strata). Further, within the RT stratum with 39 subjects/intervention arm, we can detect a difference of |33.3%| and within the chemoRT stratum with 21 subjects/intervention arm, we can detect a difference of |46.7%|. Within the smallest stratum of chemoRT and Proton RT, for example, we would have 9 patients intervention group and a comparable detectable difference of |65%|. While this detectable difference is very large, there is no reason to expect heterogeneity among the strata. However, we will examine the homogeneity of the effects across the four strata using a Gail-Simon test for qualitative interaction. Unless there is a statistically significant difference (at the 0.10 level) in this test for qualitative interaction with respect to outcome and treatment effect in the four strata, results will be combined. Unless otherwise specified, all sample size calculations were performed using PASS 2014 software, NCSS, J. Hintze, Kaysville, UT. Similar sample size considerations would apply to the end of study comparison of OM grade. We note that we evaluate the primary outcome at both the end of RT/chemoRT (Visit 9/FIV), and at the end of study visits (Post-RT Visit). These two time points will be evaluated with no adjustment for multiple analyses.

For the quantitative secondary outcomes that include changes in salivary proinflammatory cytokines, and the following clinical indicators: a) duration of and time to onset of severe OM, b) salivary hypofunction, c) average mouth and throat soreness [MTS], and d) quality of life and function as measured by selected instruments [EORTC QLQ 30 and EORTC QLQ HN35]), we can detect a difference of |.5 | standard deviations of the difference between the two intervention groups with 60 participants/group for a single measurement at a single timepoint with 2-sided alpha of 0.05 and power of 80%. These detectable differences range from |.6| standard deviations to |1.4| standard deviations as the strata size ranges from 9 to 39 per intervention group.

Details for the one planned interim analysis of the primary endpoint are provided in Section 12.4 below with the fixed planned study sample size of 120 participants.

12.3. Planned Interim Analyses

A formal interim analysis for efficacy and futility will be carried out when 50% of the participants have been accrued and would have been evaluable for the 3-month post treatment visit. With the planned accrual of 120 participants over 3.5 years, we expect this analysis to occur at approximately 2.5 years after the first patient is randomized to the study**.** The planned timing of this interim analysis is early enough to allow sufficient time to impact on the trial accrual within the time frame of this trial. That is, this planned interim analysis will occur while patients are still actively being randomized to treatment. Randomization of participants would not be halted during the review period. The table below provides the detectable differences and alpha and beta spending for the planned interim and final analysis. The boundaries for efficacy and for futility at the interim and final analyses are provided in Table 1 below that provides the detectable differences and p values.

Table 1. Detectable Differences and p values for Interim and Final Analyses to Detect a Difference in the Proportion of Patients OM Response of |0.265| on Intervention Compared with SOC (0.30) based on z test to compare proportions (Casagrande-Pike-Smith Correction], unpooled variance estimate 2-sided α =0.05; power =0.80; 60 patients/treatment group O'Brien Fleming Boundaries, Lan-DeMets Spending Function for Efficacy and Non-Binding Futility.

Table 1.

Analysis	*Number of Participants*	*Cumulative α Spent*	*Efficacy Boundary*		*Cumulative β Spent*	*Futility Boundary*	
			Difference Treat-SOC	p-val, z statistic		Difference Treat-SOC	p-val, z statistic
Interim	60	0.003	|0.39|	≤0.003	0.039	|0.046|	>0.724
Final	120	0.050	|0.183|	≤0.049	0.198	|0.183|	>0.049

[Calculations from EAST V6.4.1 Cytel, Inc.]

12.3.1. Safety Review

SAEs and related AEs will be summarized by treatment group, strata, visit and body system for all participants who received the intervention. Safety and toxicity will be summarized to compare the distributions of incidence and severity of adverse events on the two intervention groups within each treatment stratum. Frequency distributions and time to event analyses (Kaplan-Meier curves) will be used to summarize the major events of interest.

12.4. Analysis Plan

Distributions of subject and disease characteristics will be summarized by randomized intervention group (and within strata by intervention group) using descriptive statistics and graphical displays (e.g. boxplots for continuous measurements and frequency displays and contingency tables for ordinal and categorical measurements). These summaries will be provided at baseline and at each study visit, including the end of study visit (Post-RT Visit). Similar descriptive analyses will be provided for all secondary outcomes as well (changes in salivary proinflammatory cytokines, and the following clinical indicators: a) duration of and time to onset of severe OM, b) salivary hypofunction, c) average mouth and throat soreness [MTS], and d) quality of life and function as measured by selected instruments [EORTC QLQ 30 and EORTC QLQ HN35]). The focus will be to identify potential heterogeneity and imbalance of randomization into the study intervention groups.

Primary Objective. The analyses of the primary endpoint will be carried out on the binary OM severity (WHO OTS 0-2, 2-4) at the end of RT/chemoRT and at the end of study using a Cochran Mantel Haenszel test to adjust for strata if there is no evidence of heterogeneity of treatment effect. If there is heterogeneity, results will be examined separately within strata. Individual trajectories of OM severity by visit will be displayed.

Additional analyses of changes over time in OM severity (with ordinal classes) will be examined using mixed effects regression models for subject specific models that require only that data be missing at random [see for example; Applied Longitudinal Data Analysis, [35]or General Estimating Equation (GEE) population average models that require data be missing completely at random.[36] These approaches take into account repeated observations, missing data, and dropouts and loss to follow up among the subjects.

Effects of dropouts. Crossovers from intervention to standard of care group and noncompliance. The assumptions of these methods will be evaluated by comparing those subjects who dropout over time with those who do not with respect to baseline, diseases, and treatment characteristics to identify potential differential dropout rates by intervention group, stratification variables, and other factors. Further sensitivity analyses will be conducted to evaluate the effects of dropout rates that may be different between intervention and control group; crossover from intervention to control group; noncompliance on intervention and SOC arms, and missing outcome and covariate data.

Adjustments for covariates. Analyses will incorporate adjustments for baseline variables (e.g. age, gender, cancer diagnosis, tumor location, RT treatment plan and modality, RT dose and targeted sites within the oral cavity) including dental health status and other factors that may impact outcome based on the analyses that compare baseline variables between the two treatment groups. Those variables that are clinically important or that differ between the two treatment groups at baseline will be considered in these supportive analyses.

Secondary Objectives. Similar methods will be employed for the analyses of the secondary outcomes that include comparisons of changes in cytokine levels over time by treatment intervention and comparisons of changes: a) duration of and time to onset of severe OM, b) xerostomia, & c) average mouth and throat soreness (i.e. other clinical effects of the intervention), quality of life assessments, cytokines, etc.; that are evaluated over time in this study.

WHO Severity compared with CTCAE score. Data will be summarized using contingency tables for these pairwise measurements. The association between the two grading systems will be examined using nonparametric McNemar chi square tests for symmetry.

Cytokine Analysis. The focus of the analyses of cytokines is to examine the effects of the intervention compared to the control group on OM response that incorporates all the various cytokines measured over time. The strategy for the development of these prediction models for OM response will include logistic regression models (and mixed effect regression models) calibrated with cross validation or penalized logistic regression models such as LASSO, [37]and other methods[38] to identify salivary cytokines that, individually or jointly, may play a role, in addition to intervention assignment, stratification factors, and patient characteristics, in the development of OM and the severity of OM over time. The distributions of each of the cytokines will be summarized within intervention groups at baseline and over time with summary statistics and graphical displays. Bivariate scatterplots and pairwise correlation coefficients will be estimated. Preliminary analysis will compare the distributions of individual cytokines by treatment group using 2-sample 2-sided t-tests (with suitable transformation of levels if required). With a false discovery rate of 10%, and a power of 80% for each test, we can detect a true difference in cytokine level of at least 0.5 (with estimated within group standard deviations of 1.0). For a single test, the individual test alpha is 0.074; the probability of detecting all 5 tests where the true mean difference in expression > 0.5, is 0.33.

The analysis approaches described above incorporate biomarkers, demographic and clinical predictors (and potential interactions) may increase both sensitivity and specificity in OM prediction or classification accuracy. The resulting models will be evaluated using Receiver Operating Characteristic (ROC) curves and estimating increases in AUC or other appropriate statistical measures of improvement in classification. An optimal cut-off value can be selected using cross-validation by optimizing the Youden index of the ROC analysis.[38] Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Additional approaches to the development of prediction models will also be considered. Results would require validation in an independent study.

Exploratory Analysis. Progression free survival and overall survival will be plotted using Kaplan Meier curves by treatment group and randomization strata. Median times to failure will be estimated along with 95% confidence intervals if there are sufficient numbers of events.