Statistical Analysis Plan

Study Design Overview: This is a Phase III Clinical Trial. Our overall goals are to evaluate the effectiveness of nonoperative and operative interventions in patients with adult symptomatic lumbar scoliosis (ASLS) and to identify important clinical and radiographic determinants of change in patient-reported HRQOL. To accomplish these goals, a 5-year cooperative, multi-center longitudinal randomized study with a concurrent observational cohort is proposed.

> **Specific Aim #1:** Compare the outcomes of surgery and nonoperative treatment in patients aged 40 to 80 with ASLS defined as a lumbar curve with a coronal Cobb measurement of 30° or more, and either of the following: Oswestry Disability Index (ODI) score of 20 or more; or Scoliosis Research Society Quality of Life (SRS-QOL) instrument score of 4.0 or less, in the domains of pain, function and/or appearance.
> Null Hypothesis: Nonoperative and operative treatment groups will have similar outcomes at follow-up.
>
> **Specific Aim #2:** Evaluate the impact of patient factors (age, gender, socioeconomic status, education) and comorbidities [mental health, body mass index (BMI) and bone mineral density (BMD)] on adverse events and treatment outcomes for both the non-operative and operative arms. Incorporate these variables into a prediction model to help identify those patients most likely to benefit from either a surgical or non-operative approach.
> Hypothesis: Age, gender, socioeconomic status, education, and comorbidities will have an impact on the final result at 2-4 years post-intervention.

1.6 Statistical Analysis. The primary analyses will use longitudinal regression modeling (Using SAS 9.4 and the Mixed procedure for continuous and the GLIMMIX procedure for categorical outcomes), where outcomes being assessed will be change from baseline and the primary covariate will be the baseline outcome score. These longitudinal modeling techniques that incorporate various baseline covariates, such as baseline outcome, patient age, gender, and other patient- and treatment-specific factors will allow evaluating of both differences in change from baseline 1. Across time of assessment by treatment groups adjusted for baseline covariates and 2. Treatment group differences at any specified point of time.

Preliminary analyses prior to the longitudinal regression modeling will evaluate Treatment group differences at baseline in both the RCT and Observational cohorts in the RCT study, observed differences in HRQOL or personal demographics will evaluate the success of the randomization in the OBS study, observed differences in HRQOL and/or personal demographics will inform decisions regarding what covariates should be used to control for these baseline differences to allow better interpretation of longitudinal treatment differences These analyses are likely to be done using recursive partitioning rather than stepwise regression methods, which, of late, have fallen into disfavor.

Sample size calculations to ensure adequate power were conservative by focusing on the ability to detect statistically significant differences between treatment groups at any specific follow-up interval. These are essentially a priori tests of simple effects of a treatment by time interaction, which is the primary focus of the study. These analyses are considered conservative for two primary reasons.
These differences are "between subjects" effects, meaning that the added power associated with repeated assessments across time do not help to increase power
The standard assumed standard deviations in the calculations did not consider the standard deviation estimate shrinkage likely to result from the included baseline covariates (mostly baseline outcome).

1.6.1 General Analysis Strategy: Our approach to data analysis includes three consecutive stages. Most analyses will be performed using SAS software, which has a wide range of statistical routines and offers great flexibility in data manipulation.

Stage 1 – Descriptive Analysis:
Quality control will be ongoing throughout the project, as part of data management procedures that monitor data quality at each site and overall. However, prior to analysis, the data will be fully examined to confirm data quality and make final corrections. After addressing data quality, each variable will be summarized via marginal distributions as well as by relevant separate and combined study factors [(e.g., treatment, crossover status, patient demographics (age, sex, race, education)], and clinical factors [(e.g., curve type and magnitude, stability across time and outcomes (SRS-QOL, SF-12 PCS and MCS, ODI, etc.)]. Distribution summaries for continuous variables will include means, medians, interquartile ranges, minima and maxima and standard deviations; those for categorical variables will include proportions and 95% confidence intervals. Graphical methods–including histograms scatter plots, box plots, and plots of outcomes by time–are particularly important at this stage in order to understand aspects of data quality. The amount and patterns of missing data will be examined early and regularly during data collection to detect possible problems with instrument completion. Prior to analysis the amount and patterns of missing data are also important to see whether adjustments will be needed at the Inferential Analysis Stage. Those who complete data collection will be compared with those who drop out before the end of follow-up to identify potential selection biases, causing the study pool to differ from the target populations.

Stage 2 – Exploratory Analysis and Data Reduction:

a. Construction of Index Variables and Data Reduction: In some instances, data collected in the questionnaires will require pre-processing prior to data analysis. For example, the SF-12 can be split into two domains using well-defined algorithms, where each domain is scaled between 0 and 100: a Physical Construct (PCS) will reflect the patient's functional status, and a Mental Health Construct (MCS) will reflect the psychosocial aspects of the patient's condition. Similarly, the SRS-QOL is typically reduced from 30 items to key sub-scales, such as pain, function, mental health, and self-image.

b. Variable Transformations: Transformations will be employed as needed to produce variables that conform to the distributional assumptions underlying the regression models that will be used. For example, some variables may be transformed to the logarithm scale for analysis, to reduce any marked positive skew; in these cases, results will be reported on the original scale of the variable to aid in interpretation. Transformations may also be required to ensure additivity of variable effects in general linear models (for example, the outcome may be a quadratic rather than a linear function of age).

Stage 3 – Inferential Analysis:

Specific Aim 1: Compare the outcomes of surgery and non-operative treatment in patients aged 40 to 80 with ASLS defined as a lumbar curve with a coronal Cobb measurement ≥ 30° and either of the following: Oswestry (ODI) score of 20 or more, or Scoliosis Research Society Quality of Life instrument (SRS-QOL) score of 4.0 or less in the domains of pain, function and/or appearance.

a. Primary Analyses: All the data analyses required to fulfill the specific aims of the study will be accomplished using statistical methods for repeated measures and longitudinal data using the SAS system's MIXED and GLIMMIX procedures for continuous and categorical outcomes, respectively. These procedures use maximum likelihood methods to estimate variances associated with the fixed and random effects of the model. In the randomized cohort, the primary analysis will be on an intent-to-treat basis. For the observational cohort, differences between treatment groups are expected at baseline and these will be controlled for by using covariate adjustment within the longitudinal regression models. We will model each HRQOL measure as the change at follow-up from baseline at each study month (i.e., at 3, 6, 12, 18, 24, 30, 36, 42, and 48 months), adjusting for baseline level while clustering the vector of outcomes within patients. Results will provide estimated means and standard deviations for the rate of change over the 12-month and 24-month (primary endpoint) follow-up periods, (1) assuming a linear rate of change, and (2) not assuming linearity. We will also report the 3-year and 3-year changes (and 95% confidence intervals) predicted by these models. For SRS-QOL and SF-12, component domains will be analyzed both combined and separately as described by Mangione et al.

b. Control of Multiple Comparisons. A variety of outcomes will be used to measure effects within different domains of interest in this longitudinal study, all of which may be correlated. Initially, each outcome will be analyzed separately; these analyses are described in detail in the following section. In this section we present methods by which the outcomes will be examined jointly. We will apply the method developed by Bloch et al.[52] for comparing effects of two treatments on *J* multivariate outcomes, where these outcomes include SRS-QOL subscores, SF12, and ODI measures of QOL.

Let the difference between arm means for outcome *j* be given by $\Delta_j = \mu_{1j} - \mu_{2j}$, *j=1,…,J*. Changes in all the relevant outcomes will be defined so that all differences are in the same direction and we can test for a positive one-sided multivariate difference. If surgery is treatment 1, then surgery will be deemed more effective if:

Criterion 1: $\mu_{1j} > \mu_{2j}$, for some *j* (shows superiority of surgery)
Criterion 2: $\mu_{1j} > \mu_{2j} - \varepsilon_j$, for all *j* (shows non-inferiority of surgery).

(Extensive additional details available in the source article are better properties than a multivariate test based only on criterion 1. Furthermore, if the evidence for most outcomes meets the stricter first criterion, then the overall evidence (across all outcomes) may still conclude efficacy. An advantage of this approach is that it accommodates correlation among the outcomes, whereas the Bonferroni correction not included here.) Bloch et al.[52] showed that a test based on both criteria has does not. Further, it provides a global test that considers multiple dimensions of the mechanism of action of the treatment, avoiding confusion over which of several p-values to report. Finally, this method is nonparametric and thus does not require normally distributed outcomes or rely on the central limit theorem. This feature may be especially important in the current project where we rely on ordinal outcome scales.

Specific Aim 2: Evaluate the impact of patient factors (age, gender, socioeconomic status, education) and comorbidities [mental health, BMI and BMD] on adverse events and treatment outcomes for both the non-operative and operative arms and incorporate these into a predictive model to help identify those patients most likely to benefit from either a surgical or non-operative approach.

The models describing the experiences of nonoperative and operative patients, discussed under Aim 1, will be expanded to determine if the time trends depend on baseline covariates, including age group, gender, socioeconomic status, education level, and comorbidities, including BMD, BMI, and mental health status. Each baseline predictor initially will be added to the model separately, to determine its association with the outcome trend. We will then use a backward selection procedure to reduce a “full” multivariate model that includes all baseline variables that are found to be statistically significant in the univariate analyses, to a “parsimonious” model that includes only baseline variables that remained statistically significant in the multivariate analysis. For these analyses, “statistical significance” will be assessed at the 0.10 levels, in order to err on the side of including baseline characteristics that may predict outcome. Note that our purpose here is not significance testing; rather, the p-value is used to screen for covariates that are prognostic. We will summarize these findings using predicted levels of change in HRQOL generated by the models. These models will be used to characterize patients who achieve exceptionally good outcomes and those who achieve exceptionally poor outcomes. These results will be useful to clinicians in identifying patient prognoses.

1.6.2 Sample Size Requirement: The experimental design is a 2bTrt (2 between-subjects factor levels: operative and non-operative) x 9wTime (9 within-subjects or repeated factor levels: 3-, 6-, 12-, 18-, 24-, 30-, 36-, 42-, and 48-month follow-up) factorial design. For the observational cohort study, the two treatment arms are not expected to be equivalent at baseline and, therefore, the data analysis is planned as a mixed-model longitudinal analysis with the change in SRS-QOL outcome at each follow-up point relative to baseline modeled from the baseline score and selected baseline covariates across time. Although follow-up times are fixed (as specified in the design), there is likely to be some variance in actual assessment times. Therefore, SRS-QOL outcome will be modeled by specifying a random intercept and basing the estimate on the date of assessment within each fixed follow-up assessment interval relative to the distance (in days) of the actual follow-up to the center point of the follow-up interval. In this way the longitudinal model includes both fixed (treatment and assessment period) and random (time deviation of assessment within each assessment period)

effects. The MIXED procedure in SAS Version 9.2 (SAS Institute, Cary, NC) will be used to evaluate these results and will provide overall all estimates of Treatment, Assessment Time and the Treatment x Assessment Time interaction effects. For the RCT group, although randomization at baseline is likely to diminish any baseline differences between the two groups, the same analysis approach will be followed. In this case, however, the covariates will be chosen based on their ability to explain and, therefore, help to compensate for possible attrition and/or crossover effects.

Sample sizes needed to achieve power of .80 for simple effects tests treatment differences at any given follow-up interval. Needed per group: 32-37 patients.

Evaluating the statistical power for a mixed model longitudinal analysis is not straightforward. One conservative approach is to power the study with regard to a simple effects test associated with the highest order effect of interest in the study, here the Treatment x Assessment Time interaction. In this case, the sample size for the two treatment groups is estimated and, since the two treatment groups are not matched, the sample size per treatment group is likely to be larger than would be required to establish a significant change in outcome within a Treatment group across Assessment Time simple effects test (e.g., difference in SRS scores from baseline to 2-year follow-up). Accordingly, this study was powered with respect to the simple effects tests associated with a significant Treatment x Assessment Time interaction at the primary 2-year endpoint. This power analysis is informed by the data already gathered by the SDRG with the mean and standard error estimated adjusted for SRS baseline and patient age gender and Cobb angle. ***Table 2 summarizes estimated sample sizes per treatment needed to achieve power of .80 for 2-tailed testing with type I error rate set at .05 under a variety of effect sizes, based on mean and standard deviation estimates***. The two rows highlighted were judged to reflect the most likely scenarios and, overall, the sample size of 30 per group in the RCT was considered a reasonable choice for achieving sufficient power to detect a real difference between the two treatment groups at 2-years follow-up.

Table 2

Estimated Means		Common SD	Effect Size	N per Group
Trt 1	Trt 2			
0.3	0.8	0.6	0.833	24
0.4	0.8	0.6	0.667	37
0.5	0.8	0.6	0.500	64
0.3	0.8	0.7	0.714	32
0.4	0.8	0.7	0.571	50
0.5	0.8	0.7	0.429	87

1.6.3 Participants Available for Recruitment and Power. From the beginning of 2004 through the end of 2006, the five participating sites enrolled 155 non-operative patients and 141 operative patients meeting the criteria for diagnoses, symptoms and age required for this study (Table 3, next page). These preliminary data show a consistent recruitment level that should allow us to enroll our estimated 300 total subjects. The SDSG group estimates that over 90% of patients are willing to participate in a clinical trial, with 30% of those willing to randomize. If, over the course of 3 years, 300 patients are eligible, then an estimated 270 (90% of 300) would be willing to participate, with 82 of these (30% of 270) willing to randomize. With 1:1 randomization to operative and non-operative treatments, at the end of 3 years, 41 patients would be randomized into each of the operative and non-operative arms. With 20% crossover in the non-operative arm, there would be three treatment groups: 41 in operative, 33 in non-operative, and 8 cross-overs from operative to non-operative, although in an intent-to-treat analysis all of these crossovers would be retained in their initially assigned group. Assuming a 10% loss to follow-up by 2-years, these numbers are reduced to 36, 30, and 7 for the operative, non-operative and crossover groups, respectively. Based on Table 2 (above), as shown in the “boxed” rows, the 80% power to detect effects sizes at or above 0.714 and 0.814 would still apply to the RCT. Further, within the longitudinal analysis structure, patient data remains to inform the model for every follow-up assessment

gathered, and therefore, a patient's loss to follow-up does not negate the information that they provided up to that point, thus providing additional precision to estimates that will boost statistical power.

For the observational cohort, applying the patterns suggested above (270 willing to participate) it follows that 188 would be willing to participate in the preference trial. If we assume a 1:1 surgical preference, which is not unreasonable given the SDSG preliminary data, this would result in 94 non-operative cases and 94 operative cases. Assuming a 10% crossover to surgery from non-operative care, this would result in 85 non-operative and 103 operative cases and, with 10% lost to follow-up, the final expected cell counts would be 77 and 93. In these circumstances, a two-group t-test with a 0.050 two-sided significance level will have 80% power to detect a difference in means of 0.175, assuming that the common standard deviation is 0.400, when the sample sizes in the two groups are 77 and 93, respectively (a total sample size of 170). An observational cohort of this size will supplement the RCT very well in that longitudinal analyses of these data will support evaluation of a number of covariates. Furthermore, similarities and differences in results between the RCT and observational study will provide some foundation for establishing (or not) the generalizability of the RCT results to the clinical case where randomization to treatment is not practiced.

Table 3: Breakdown of Enrollment from Five Centers, 2004 through 2006, According to Age Group and Intervention Arm

	Nonoperative (N=155)						**Operative (N=141)**					
	2004 (N=59)		**2005 (N=43)**		**2006 (N=53)**		**2004 (N=41)**		**2005 (N=41)**		**2006 (N=59)**	
	40-59	60-80	40-59	60-80	40-59	60-80	40-59	60-80	40-59	60-80	40-59	60-80
Washington University, St. Louis, MO	9	5	0	3	3	4	14	2	14	10	16	8
Emory Spine Center, Atlanta, GA	13	6	4	7	8	7	13	3	4	4	9	8
Leatherman Spine Institute, Louisville, KY	6	9	4	4	8	8	0	5	4	1	5	2
Univ. of Virginia, Charlottesville, VA	1	6	6	12	1	8	1	2	0	3	3	2
New York University, New York, NY	2	2	1	2	2	4	1	0	1	0	1	5
TOTAL	31	28	15	28	22	31	29	12	23	18	34	25

1.6.4 Stopping Rule protocol:

From the study power analysis, power of 80% to detect an effect size between .667 - .714 with type I error rate set at .05 and 2-tailed testing requires a sample size of 32 – 37 patients per treatment group. We plan (are budgeted) to enroll 300 patients overall and expect about 100 (1/3) of these patients to be randomized. For the purposes of a stopping rule, N counts needed to establish power greater than 70 are more reasonable given stopping a study for futility is quite different than stopping rules associated with safety considerations. Power of 70% with an effect size of .667 requires 29 patients per group. A calculator estimating the probability of obtaining at least 58 patients from a sample of 300 enrolled patients under various recruitment and attrition scenarios

Based on this calculator results, futility risk (i.e., the inability to recruit at least 58 patients into the RCT) is quite high if, after 100 recruited patients, RCT recruitment is at no more than 20%, even if attrition rates are as low as 10%. When RCT recruitment rates are up to 25%, futility risks are relatively low when attrition rates as also low (≤ 15%) but otherwise probability of successful recruitment are low. When RCT recruitment rates reach 30% for the first 100 enrollees, the probabilities of successful recruiting at least 58 patients is quite good, as long at attrition rates remain no greater than 30%.

In sum, if we calculate the recruitment and attrition rates in RCT study after 100 patients have been enrolled we can estimate the likelihood of attaining an RCT sample size of at least 58, which would result in power in the 70% range. If likelihoods of attaining sufficient patients are low based on the first 100 patients enrolled, the choices are: 1) stop the RCT enrollment; or 2) close the preference trial when 200 patient shave been enrolled and continue to recruit into the RCT. The decision to continue enrolling the RCT after the preference trial has closed is predicated on assumptions that a) patients might be more willing to enroll into the RCT given there is no other study choice; and b) the nature of the patients enrolling into the RCT after the preference trial has closed are from the same population. Therefore, if the second choice is taken, a subsequent futility stopping rule for the RCT would be in force. This stopping rule will include two arms: 1) the same as the first stopping rule regarding recruitment and attrition and 2) an evaluation of the similarity of the RCT patients enrolling prior to and following the closing of the preference trial.

Addendum 8/29/12: (modified and edited 6/19/13) The primary analysis plan for this study is to first evaluate the RCT and OBS studies separately using mixed model longitudinal regression analysis using an intent-to-treat approach. As was done in the SPORT analysis, assessment Intervals are treated as a fixed effect, meaning with each interval developed based on the date of intervention. Data gathered prior to date of intervention (DOI) will be considered baseline and fixed data points gathered after DOI will be classified according the form designation (e.g., 3mo, 6mo). Data points gathered after DOI that were not fixed points (e.g., complications, severe adverse events) will be associated with a fixed interval based on the length of time from the data point and DOI. A second random effects variable (Delta) will be created within each interval that reflects the distance (no. of days) between the actual observed point and the center of the fixed interval that this point falls within. Thus, patients may have multiple data points within a fixed interval and the distance orTime of data collection relative to DOI within each fixed interval will be included in the model. In addition, a Delta random variable, defined as the distance between the center of the fixed interval and the Time of data collect will, also be created. The primary analysis will regress the change in outcome (eg for the pcs outcome d_pcs) on baseline pcs (b_pcs), treatment, interval, site and patient age and sex,

For the RCT cohort, the base model for PCS would be

d_pcs = b_pcs treatment | interval | time site age sex

A secondary analysis will replace time with the Delta variable with the model forcing the Delta covariate to have a value of 0 – thereby adjusting the results to reflect the expected value for the outcome as if the outcome was assessed at the exact center of the fixed interval.

d_pcs = b_pcs treatment | interval delta site age sex delta / delta= 0

Where:

T_DOI = Date dependent variable was collected – DOI WITH negative T_DOI values indicating outcomes being assessed prior to treatment and positive T_DOI values indicating the time after DOI that the outcome was evaluated. Note that T_DOI values are nested within d_pcs
= -pcs-b_pcs (a positive change would reflect improvement)

b_pcs = baseline pcs

Treatment = is a fixed effect with 2 levels the operative or non-operative care of the patient

Interval = is a fixed effect with X levels (depending on the outcome) indicating the follow-up times for collected patient outcomes

Time = The number of days between the date of assessment within an interval and the date of Intervention

site = the study site enrolling the patient (note site interactions may not be estimable in which case the interaction effects with site would be removed from the model

age = a preplanned covariate

sex = a preplanned covariate

delta = a random effect reflecting the distance (in days) of the observed score in an interval from the center of the interval

Comparing the differences in these two modeling approaches may be useful for evaluating the relative merits of the two analysis approaches.,
For predictive modeling crossover will not be considered and as "Intent to treat" modeling – where the assigned treatment (in the RCT) or the selected treatment (in the OS) cohorts will be retained throughout the analysis.
Additional baseline covariates may be added to this model depending on the results of preliminary comparison of baseline differences between treatment groups and recursive partitioning methods will explore additional possible baseline factors and interactions among these factors to detect patient profiles that vary across treatment assignment. In addition, significant relationships between baseline variables and crossover status and loss to follow-up will also be useful for identifying covariates.
For the OBS group, where baseline differences between the two treatment groups are expected, the same basic model will be used and the same approach for adding covariates to the model based baseline differences between treatment groups, crossover status and loss to follow-up will be used.

The primary hypothesis of interest concerns differences between treatment outcomes at each time interval and differential change in outcomes within treatment across time intervals. Thus, preplanned follow-up simple effects tests associated with the treatment x interval interaction will be explored even if the interaction itself is not statistically significant (this is what preplanned means).

These simple effects tests will evaluate:

1. Differences in effect size (e.g. d_pcs) at each interval between the two treatment groups (e.g., are the effect sizes between treatment groups at 2 years significantly different?);
2. Differences in effect sizes within each treatment group across the intervals (e.g. is the mean effect size for surgery at 3-months significantly different from the mean 6-month effect size?).
3. Differences in effect size change from one interval to the next between the treatment groups (e.g., are the slopes from 1 to 2 year follow-up significantly different across the two treatment groups?)

These analyses will

a. Calculate a clinical trial p-value for each outcome at the 12- and 24-month time points
b. Provide mean scores for each continuous outcome across all intervals, with conditional simple effects tests allowing for differential trajectories across time for the treatment groups.

Proc mixed in SAS version 12.1 64-bit running under the Windows 7 Ultimate 64-bit operating system will be used to analyze these data.

Evaluating Crossover. The exact approaches that will be used to handle crossover effects will depend on the magnitude and pattern of crossovers. Crossover in both the RCT and OBS studies will occur, and probably in both directions (non-op to surgery and surgery to non-op – meaning patients randomized or opting for surgery will never have the surgery).

Our approach for evaluating the effect of crossover would be to create a crossover variable to be included in the model that would have two possible values (0 – indicating no crossover; 1 – indicating crossover). By continuing with an Intent to treat philosophy for analysis, adding crossover status to the longitudinal regression modeling (along with interactions of crossover with treatment, interval and site) will provide a means for directly evaluating the effects that crossover will have in the model. Note that these models will be explanatory rather than predictive since crossover status will remain unknown until it occurs and, thus, is not a useful predictor. Including a crossover*treatment interaction will indicate (within each study cohort) who did and did not crossover and what the nature of the crossover was. Significant effects involving the crossover variables relative to outcomes and baseline characteristics, and interactions between the crossover variable and other

study factors (Treatment, Interval, Site) would be useful in determining what other covariates might be needed in the model to most accurately evaluate treatment efficacy across time. Our group at Dartmouth is considering using this approach for examining the effects of crossover in the SPORT trial as well.

Treadmill evaluation. Again, longitudinal mixed model regression analysis will be used to evaluate the treadmill evaluation. However, since crossover to surgery will trigger a "new" baseline treadmill test, these patients will be repeated in the analysis with their 2nd baseline considered their follow-up treadmill test relative to the 1st baseline score and their final treadmill test 2 years following their 2nd baseline treadmill test will be considered their follow-up treadmill test relative to their 2nd baseline treadmill test.

Treadmill = treatment | interval| time site age sex
Treadmill = treatment | T_DOI(interval) site age sex

Patients who crossed over to surgery will be provided with a second DOI value (DOI_cNS) that will be used in conjunction with the date of the follow-up treadmill test to define the delta associated with the 2 year interval data form used to gather the follow-up treadmill score.

Addendum 1/17/13: Sample Size Calculation.
A sample size of 82 patients (41 per treatment arm) was estimated to provide 80% power with an SRS-22 effect size of ≥0.714, 10% loss-to-follow-up at 2 years postoperative, and 20% crossover from nonoperative to operative intervention, based upon a mixed-model longitudinal analysis accounting for repeated measures at follow-up. A priori sample size estimations require estimates of outcome variances that may be inaccurate in the actual population studied. To account for this a bootstrap sample size estimation using data from the first 26 patients randomized was performed to provide a more accurate assessment of the required sample size needed to address the original trial aims, to include both the SRS-QOL and ODI.

Sample Size Estimation

	ODI	SRS	
Interval	Power	Power	n/GRP
3 months	0.0390	0.6770	18
6 months	0.7270	0.9980	
9 months	0.5600	1.0000	
12 months	0.8170	0.9950	

These results suggested that >80% power could be achieved to address Aim #1 with 36 randomized patients (18 patients per group).

Enrollment in the Observational cohort closed May 2013. Enrollment in the randomized cohort closed July 2014 with 63 participants, 70% of the original enrollment goal, this enrollment goal was felt, based on the above sample size analyses, to allow for sufficient power making allowances for loss to follow-up and cross-over over time.