Cytel Consulting

Experience • Expertise • Innovation

Cytel Consulting Document Addressing

Statistical Analysis Plan for the PORSCHE Trial

PROTOCOL No.: 16.21.CLI

for

Nestlé Health Sciences

June 22, 2018

PREPARED FOR:
Natalia Muehlemann
Global Business Manager - Medical Devices
Nestlé Health Sciences
Phone: + 41 (0) 21 924 7984
Email: Natalia.Muehlemann@nestle.com

PREPARED BY:
Rajat Mukherjee, PhD and Jian Wang, MS
Cytel Inc.,
675 Massachusetts Ave.
Cambridge, MA 02139
Email: *rajat.mukherjee@cytel.com*
Phone: +34 660 010 852

Statistical Analysis Plan for the PORSCHE Trial PROTOCOL No.: 16.21.CLI

Contents

List of Tables

List of Figures

1 Introduction

This document describes the Statistical Analysis Plan (SAP) for the final analysis as well as for multiple interim analysis for the PORSCHE trial: A prospective multi-center single-blinded study comparing the performance of the Dysphagia Detection System (DDS) in detecting impaired swallowing safety and efficiency as compared to the clinical reference method - videofluoroscopic swallowing study (VFSS).

The core component of the DDS is a statistical classifier algorithm that is able to predict the probability that a swallow was normal or with impaired safety and/or impaired efficiency based on the swallow signals captured using a dual-axis accelerometer sensor. The statistical classifier is built on a training data set collected from 305 subjects in an exploratory (phase-0) trial. Details can be found in the statistical report for this trial.

The main objective of the PORSCHE trial is the validation of the DDS.

This SAP has been developed for the PORSCHE trial based on Protocol 16.21.CLI (Amendment No. 2 dated 25 September 2017). The SAP describes the statistical methods to be used for the analysis and reporting of data collected during the conduct of Protocol 16.21.CLI, and has been developed and finalized prior to database lock and any unblinding of the clinical database for Study 16.21.CLI. This is an integrated SAP which includes the statistical analysis plan for interim analyses as well as the final analysis. If additional analyses are required to supplement the planned analyses described in the SAP, they will be identified in the Clinical Study Report (CSR).

This SAP is being written with consideration of the recommendations outlined in the International Conference on Harmonisation (ICH) E9 Guideline entitled Guidance for Industry: Statistical Principles for Clinical Trials and the most recent ICH E3 Guideline entitled Guidance for Industry: Structure and Content of Clinical Study Reports.

2 Study objectives and trial design

2.1 Study Objectives

The primary objective of the PORSCHE trial is to validate the DDS for detecting impaired swallow safety using thin-Barium (Thin-BA) swallows (as a proxy to water swallows) in terms of sensitivity and specificity with respect to the gold-standard results obtained using simultaneous VFSS. Barium contrast is needed for conducting the VFSS.

Secondary objectives are to validate the DDS for detecting impaired swallow safety for thicker consistencies: mildly thick Barium (Mild-BA) and moderately thick Barium (Mod-BA) and also to validate the DDS for detecting impaired swallow efficiency using thin, mildly thick and moderately thick Barium swallow. The validation is again to be carried

out in terms of sensitivity and specificity with respect to the simultaneously carried out VFSS results.

2.2 Study Procedure

Each participating subject goes through a series of swallows using 5 Thin-BA boluses, 4 Mild-BA boluses and 4 Mod-BA boluses. During each swallow, signals are captured using a dial-axis accelerometer which is then fed into the DDS device for classification as impaired or normal. Simultaneously, for each swallow, the swallowing process is also captured using VFSS which is then assessed by a central VFSS reader. The VFSS reader is blinded to the DDS results. Only 4 out of the 5 Thin-BA swallows and 3 out of the 4 Mild-BA and Mod-BA swallows will be used for statistical analysis. The extra swallow will be used only in cases where one of the first 4 in case of Thin-BA and one of the first 3 in case of Mild-BA and Mod-BA have missing VFSS label. Missing VFSS labels can arise due to pure video quality, disagreement between the two independent central VFSS readers or other technical issues.

2.3 Study Design

The study is designed as operationally seamless to facilitate the updating of the thresholds on the Receiver Operating Characteristic (ROC) curves for swallowing safety if required and to validate the DDS classifier with a fixed threshold using an independent validation set. The threshold updating will carried out separately for each of the three consistencies.

PORSCHE trial will start with the frozen classifiers based on fixed thresholds derived using ROC analysis from the completed exploratory trial (Protocol 11.48.CLI). The trial will initially start as a 3-look group sequential design (GSD). At the first interim, sensitivities and specificities will be calculated based on the fixed thresholds. If the Area under the ROC curves at the first interim is above 75% but the corresponding sensitivity and/or specificity are low then the threshold will be revised based on the first interim data. In this case, the validation will exclude the data used for the first interim, i.e. the validation trial will start with the first patient enrolled after the first interim analysis data cut-off date and will be carried out as a 2-look GSD. In case the fixed thresholds from the completed exploratory trial data are not revised using the first interim data, the trial will continue as planned.

The study design in summarized in Figure 2-1.

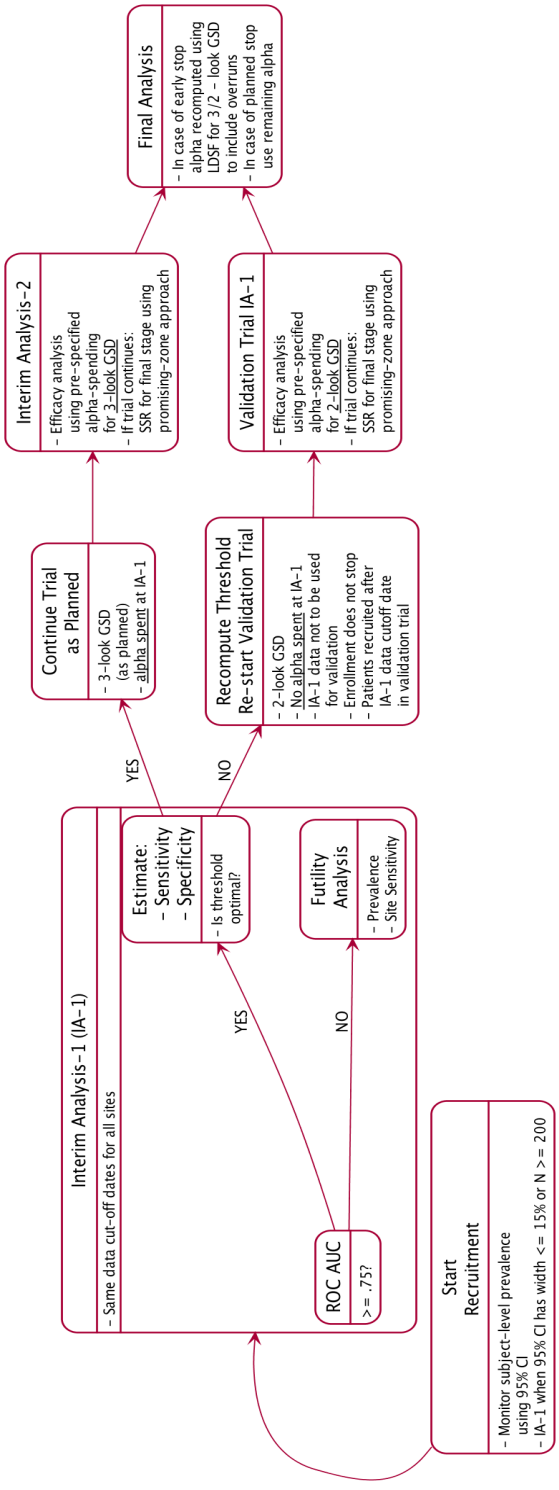

Figure 2-1: Trial Design

2.4 Efficacy Assessments

2.4.1 Primary Efficacy Endpoint

The primary efficacy of the DDS will be measured as the sensitivity and specificity obtained from comparing the DDS predicted swallow safety outcome at the subject level (from up to 4 boluses) with the clinical reference standard VFSS swallow safety outcome (binary) for THIN-Ba boluses.

Swallowing safety describes risk of penetration-aspiration which describes impaired airway protection. The impaired swallowing safety is defined as PAS ≥ 3 as determined by VFSS. (The PAS - Pentration Aspiration Scale – is provided in the Appendix B of the Protocol)

2.4.2 Secondary Efficacy Endpoints

At the final analysis, if the primary endpoint meets statistical significance then formal testing and analysis for the following secondary endpoints will be carried out in a hierarchical fashion in the specified order:

1. Sensitivity and specificity of detecting impaired swallow safety using MILD-Ba boluses (up to 3 boluses per subject).
2. The sensitivity and specificity of detecting impaired swallow safety using MOD-Ba boluses (up to 3 boluses per subject).
3. The DDS algorithm is also capable of detecting impaired swallow efficiency. The sensitivity and specificity of detecting impaired swallow efficiency using THIN-Ba boluses (up to 4 boluses per subject).
4. Sensitivity and specificity of detecting impaired swallow efficiency using MILD-Ba boluses (up to 3 boluses per subject).
5. The sensitivity and specificity of detecting impaired swallow efficiency using MOD-Ba boluses (up to 3 boluses per subject).

The success criteria for the primary and each of the secondary endpoints will be the same as specified in Section 3.1.

2.4.3 Exploratory Endpoints

The analysis on the following tertiary endpoints will be considered as exploratory:

- Prevalence of boluses resulting in "grey" outcomes of the classifier for safety and efficiency of the swallows on thin (THIN-Ba) and thickened stimuli (MILD-Ba and MOD-Ba)

- Prevalence of impaired swallowing safety and efficiency of swallows of thin (THIN-Ba) and thickened stimuli (MILD-Ba and MOD-Ba) as determined by VFSS at bolus and patient levels.
- Positive and Negative predictive value (PPV and NPV) of detecting swallow safety and efficiency problems using DDS and using THIN-Ba, MILD-Ba and MOD-Ba stimuli.
- Prevalence of impaired swallowing safety and efficiency of the swallows of thin (THIN-Ba) and thickened stimuli as determined by VFSS at bolus and patient level by the following predetermined subgroups: stroke, other neurological diseases, other patients.
- Timing and Outcome of the dysphagia screening as per usual care protocol of the study site
- Timing and Outcome of the dysphagia Clinical Swallow Assessment (CSE) by speech language pathologist (SLP) where applicable as per usual care protocol of the study site
- DDS accuracy in terms of AUC by sub-group (Stroke, Other Neurological Diseases and Others) for swallow safety and efficiency problems for all consistencies (THIN-Ba, MILD-Ba, MOD-Ba).
- Impact (or association of) of VFSS results on Nutritional Management.

2.5 Safety Endpoints

All adverse events (AEs) will be observed from the time the consent form is signed through the exit of the subject from the study.

3 Statistical methodology

3.1 Success Criteria and Hypothesis Testing

The success criteria for the DDS device validation trial is that the sensitivity is greater than 80% with a specificity greater than 50%. The sensitivity and the specificity will be estimated using an independent validation dataset. The following one-sided hypothesis testing will be carried out:

$$\begin{aligned} H_0 &: H_0^{se} \cup H_0^{sp} \text{ vs.} \\ H_A &: H_A^{se} \cap H_A^{sp}, \end{aligned} \tag{1}$$

where,

$$\begin{aligned} H_0^{se} &: \text{Sensitivity} \leq .8, \\ H_A^{se} &: \text{Sensitivity} > .8, \\ H_0^{sp} &: \text{Specificity} \leq .5, \text{ and} \\ H_A^{sp} &: \text{Specificity} > .5. \end{aligned}$$

Each of H_0^{se} vs. H_A^{se} and H_0^{sp} vs. H_A^{sp} will be tested at one-sided 2.5% level of significance which will (conservatively) preserve the over-all type-I error for testing H_0 vs. H_A.

The same hypothesis test as in (1) will be applied to the secondary endpoints. Hypothesis will be carried out in a hierarchical fashion to ensure that the type-I error is controlled at the one-sided 2.5% level.

3.2 Determination of Sample Size

The power and the sample size to test the hypothesis depends on several unknown parameters:

1. Subject level prevalence;
2. The prevalence of impaired swallows conditional on subject having Dysphagia;
3. The true sensitivity and specificity of the DDS device in detecting swallow impairment; and
4. The optimum threshold on the ROC curve.

Based on estimates from the exploratory study data, if the DDS is assumed to have a sensitivity and specificity of at least 86% and 60% respectively and approximately 35% patients show impaired safety, then a power of 90% to test H_0 vs. H_A (from equation (1)) can be achieved with 700 to 800 subjects under a fixed design. Considering that data from additional 100 to 200 subjects may be used for threshold calibration, the starting sample size is 900 subjects. A sample size re-estimation will be carried out at the last interim analysis. The maximum sample size is set to 1300 (see Table 5-2).

3.3 Group Sequential Design (GSD)

The clinical trial will initially start as a 3-look GSD. Alpha-spending will be calculated using Lan-DeMets spending function with O'Brien-Fleming parameter (LDOF) [2]. At the first interim analysis (IA-1) the threshold on the ROC curve may be re-computed using the ROC curve generated using the IA-1 data, in which case, the validation trial would start afresh following IA-1 using a 2-look GSD. Data included in the IA-1 would no longer be used for the validation phase. If the threshold is not re-computed at the first interim and

the efficacy boundary has not been crossed then the trial continues as planned as a 3-look GSD.

3.4 Timing of Interim Analysis-1

The overall subject-level prevalence of impaired safety and efficiency as determined by the central lab VFS reading will be continuously monitored in a blinded fashion, i.e. the study staff will be blinded to the DDS results. The first interim analysis will be performed when the prevalence can be estimated using a 95% confidence interval of width no more than 15%. Thus if the true (subject-level) prevalence is 20% then the interim analysis may be carried out after 110 subjects. On the other hand if the prevalence is around 50% then around 170 subjects would be needed. If the monitoring is carried out after every 50 subjects have completed the study, the first interim will be carried out around completion by 200 subjects.

3.5 Group Sequential Design and Sample Size Re-estimation Details

Due to uncertainties about the population parameters and their influence on the power, an adaptive sample size re-estimation (SSR) is proposed. This is after the first interim the trial continues as a 3-look GSD with out changing the threshold then a SSR would be done at the second interim analysis while if the threshold is changed at the first interim then the SSR will be carried out at the only interim analysis time for the freshly started validation trial post threshold re-calibration. The following adaptation rules will be applied:

1. If the threshold is changed based on the first interim look, then a new appropriately powered (80%-90%) 2-look group sequential design trial will be started. The powering will be based on the sensitivity and specificity estimates obtained at the first interim. The interim and the final analysis for this new trial would not include the data from subjects included in the first interim analysis.

2. At the interim of the new validation trial, unless the interim Z statistic crossed the efficacy boundary, a sample size re-estimation based on conditional power being in the promising zone [1] will be carried out. Conditional power (CP) will be calculated both for sensitivity as well as specificity. Promising zone is defined as both the CPs are between 50% and 90%. Futility may also be decided in case of low conditional power for either sensitivity or specificity.

3. The increase will be determined equating CP to the desired power of 90%. This will be bracketed by a minimum increase of 50 subjects and max increase defined by the maximum sample size of 1300.

4. In case the threshold is not changed at the first interim, the trial continues as a 3-look GSD and the SSR adaptation rules in 2-3 above will apply at the second interim look.

3.6 Final Analysis in case of Early Stopping

If the decision is taken to terminate the study at one of the two planned interim analysis the study will enter a closeout phase. During this phase, which extends from the soft database lock prior to the interim analysis iDMC meeting until the final database lock, additional data will arrive. The final analysis, being based on the totality of evidence, will include these additional data, resulting in an increase in the information fraction from what it was at the time of the interim analysis. The significance level for the formal hypothesis testing at the final analysis will then be recomputed from the LDOF spending function at the new information fraction. Details of this will be provided in the SAP which will be finalized and issued before the first interim analysis.

3.7 Efficacy Analysis

Efficacy of the DDS screening device will be established if it's accuracy in terms of sensitivity and specificity with respect to VFSS to detect swallowing safety impairment are high, in particular, sensitivity has to be higher than 80% while the specificity has to be higher than 50%. The hypothesis test in equation (1) in Section 3.1 is to be tested for the primary endpoint hypothesis regarding the sensitivity and specificity using Thin-BA boluses as well as for secondary endpoints listed in Section 2.4.2. The secondary endpoint s will be tested in a hierarchical fashion if and only if the primary endpoint meets statistical significance. The hypothesis testing for primary and secondary endpoint will be carried out using the following test statistics

$$\begin{aligned} Z_{se} &= \sqrt{N_d}\frac{\widehat{SE} - SE_0}{\sqrt{SE_0(1-SE_0)}}; \text{ and} \\ Z_{sp} &= \sqrt{N_h}\frac{\widehat{SP} - SP_0}{\sqrt{SP_0(1-SP_0)}}, \end{aligned}$$

where N_d and N_h are the number of diseased and healthy subjects in the analysis dataset, $SE_0 = .8$ and $SP_0 = .5$. N_d and N_h will be determined solely based on the VFSS (gold-standard) results.

At each interim the above test statistics for sensitivity and specificity will be computed and the compared to the group-sequential boundary derived using the LDOF alpha-spending function separately for sensitivity and specificity. Thus the primary efficacy hypothesis will meet statistical significance if both Z_{se} and Z_{sp} cross their corresponding efficacy boundaries. If the primary endpoint sensitivity and specificity cross their corresponding boundaries then the secondary endpoints using the same test-statistics will be tested in the pre-specified hierarchy specified in Section 2.4.2. Note that group sequential boundaries are non-binding.

3.8 Safety Analysis

All adverse events (AE) will be reported using System Organ Class (SOC) classification. The proportion of subjects experience each AE along with the mean(SD) duration will also be reported. The reporting may be done stratified by disease group and/or site if required.

3.9 Analysis Sets

3.9.1 Full Analysis Set (FAS)

The FAS is based on a modified-intent-to-treat ppulation which is defines as all subjects with a signed informed consent and at least one analyzable swallow with the Thin-BA boluses using the VFSS.

3.9.2 Per Protocol Set (PPS)

PPS includes all subjects in the FAS except for those who are excluded because of major protocol deviations. Final determinations of the major protocol deviations will be made and documented separately prior to database lock.

3.9.3 Safety Analysis Set

This consists of all subjects who have completed at least one swallow.

3.10 Statistical Considerations

3.10.1 General Statistical Considerations

All statistical analyses and data summaries are to be generated according to this SAP. Any deviations from this SAP will be documented in the Statistical Report and the CSR.

All statistical tests for the efficacy outcomes will be performed at one-sided 2.5% significance level, unless otherwise stated.

Summary statistics will be presented for the overall population as well as stratified by disease group and by site where appropriate.

3.10.2 Multiplicity

Along with the primary endpoint hypothesis (Sections 2.4.1 and 3.1) there are a total of 5 secondary endpoint hypothesis that will be tested. Alpha-inflation due to multiple

hypothesis testing is controlled by testing the secondary endpoints in a pre-specified hierarchy (Section 2.4.2) if and only if the primary endpoint hypothesis test meets statistical significance.

There are multiple interim looks proposed in the study design. Alpha-inflation here is controlled using the LDOF alpha-spending [2] function that determines how much of the total 2.5% alpha-level is to be spent at each interim look depending on the information fraction to be used at that particular interim. Thus at each interim analysis the alpha-level will be obtained using the LDOF spending function and this alpha-level will be used for the primary endpoint hypothesis testing as well the hierarchical testing of the secondary endpoints.

3.10.3 Missing Data

Missing data for the primary and secondary analysis can arise from different sources:

1. VFSS label: VFSS labels could be missing due to technical issues with the video and/or disagreement between the two independent central readers even after following procedures to resolve disgareements. Imputation of missing VFSS data for the primary and the secondary analysis will not be performed. The number of boluses with missing VFSS labels will be reported. Imputations considering "worst" and "best" case scenarios will be performed as supportive sensitivity analysis. For example, if the VFSS label is missing on one of the swallows then the worst case scenario will be to impute 1 (impaired) and the best will be to impute 0 (normal). In order to minimize missng VFSS labels, one extra swallow per consistency has been built into the protocol.
2. DSS Predicted Probability: The DSS outputs a predicted probability of impairment based on the input accelerometry signal. Due to technical reasons, the signal-t0-noise-ratio (SNR) on these signals may be unusually high and in this case the DSS output may be missing. Again imputation of these missing values for the primary and secondary analysis will not be performed, rather the number of missing values will be reported. Sensitivity analysis using "worst" and "best" case scenarios will also be performed.

3.11 Interim Analysis Plans

The efficacy analysis steps at each of the interim looks are specified below. The steps are also summarized in Appendix 5.3

3.11.1 Interim Analysis - 1 (IA-1)

The main objective of IA-1 is to determine if the fixed threshold (fixed using phase-0 data) applied to the ROC curve obtained using IA-1 data is still optimal or whether a new threshold obtained from the IA-1 data should be used moving forward. In case the threshold is changed then the validation trial re-starts while excluding the IA-1 data for testing validation related hypotheses.

The Analysis plan for IA-1 are listed as follows:

1. Compute AUC for the ROC curve based on IA-1 data
2. If AUC $< .75$, consider futility analysis. Note that this futility rule is non-binding.
3. Otherwise, calculate sensitivity, specificity both with the fixed threshold from phase-0 data as well as the optimal threshold on the IA-1 ROC curve.
4. The benefit versus the cost of changing the threshold will be analyzed. This will be done by comparing the conditional power at IA-1 under the old threshold and a total sample size of $N = 900$ ($CP_{old,900}$) with the power of a restarted trial using the new threshold and sensitivity-specificity estimates obtained at IA-1 using this new threshold along with the prevalence estimate obtained at IA-1. If the power of the restarted trial with $N = 900 - N_1$ is greater than $CP_{old,900}$ then the benefit of changing the threshold is clear. Here N_1 is the number of subjects included in the IA-1 analysis. If this is not the case then the power of the restarted trial with $N = 900$ ($P_{new,900}$) will be compared to $CP_{old,900}$. Since conditional power and power will be calculated for testing both sensitivity and specificity for the primary endpoint the formal guideline (established using simulations) is defined in terms of the product of the two as follows:

   If $P_{new,900-N_1,SE} \times P_{new,900-N_1,SP} > CP_{old,900,SE} \times CP_{old,900,SP}$ then change threshold; OR
   If $P_{new,900,SE} \times P_{new,900,SP} > CP_{old,900,SE} \times CP_{old,900,SP} + 0.025$ then change threshold.

   The above statistical logic offers a guideline for changing the threshold, however clinical justification may overrule the statistical guideline in this matter.
5. If the optimal threshold on the IA-1 ROC curve does not provide enough improvement in accuracies for the primary and secondary endpoints compared to the fixed thresholds then formal hypothesis test for the primary and secondary efficacy in the pre-specified order could be carried out at IA-1 and alpha is spent according to the LDOF spending function. If efficacy bounds are crossed for the primary and the first two secondary endpoints then the trial may be stopped early. In this case: stop enrollment and plan for the final analysis with the overruns.
6. If the hypothesis test results do not cross efficacy bounds then the trial continues as planned and moves to the next interim analysis.

As mentioned above, if the threshold is not updated using the IA-1 data then the trial continues to the second interim analysis (IA-2) or the first interim analysis for the newly started validation trial (IA-1^n).

3.11.2 IA-2 (or IA-1^n)

1. Test primary and secondary efficacy hypotheses in the pre-specified hierarchy. If efficacy boundaries for the primary and the first two secondary endpoints are crossed then the trial may be stopped early and the final analysis is planned with the overruns
2. If the efficacy boundaries for the primary endpoint and the first two secondary endpoints (i.e. for sensitivities and specificities for detecting impaired safety with Thin-BA, Mild-BA and Moderate-BA) are not crossed then consider sample size re-estimation (SSR) using the promising zone approach (Mehta and Pocock, 2011, Statistics in Medicine).
    - Calculate conditional power (CP) for sensitivities and specificities for the primary efficacy (Thin-safety). If $0.5 \leq CP \leq 0.9$ then increase sample size to meet $CP = 0.9$ up to a maximum of 1300 subjects overall (including the IA-1 subjects)
    - If $CP < 0.5$ or $CP > 0.9$ then the trial continues as planned.
    - If $CP < 0.2$ then consider futility

3.11.3 Interim Safety Analysis

Although the DDS device has been deemed Non-Significant Risk by the CDRH, FDA, the iDMC will be decide on the safety of the device and the trial based on the adverse events data presented at the interim analysis meeting by the independent statistical center. The iDMC may recommend stopping the trial if they have safety concerns about the device or the trial.

3.11.4 Interim Monitoring

Prevalence of impaired swallow safety and efficiency using different consistency swallows is a key parameter in powering of this trial. The assumed prevalence during the design-phase is between 30% - 40%. Observed prevalence may in fact be out of this range and can have serious impact on the power of the trial. For this reason, prevalence of impaired swallow safety and efficiency will be monitored in a blinded fashion throughout the trial. Furthermore, due to technical problems, it is foreseen that a small percentage of videos may not be readable by the core lab readers. Again this would be monitored in a blinded fashion.

For the purpose of blinded monitoring, only the data from the central VFSS laboratory, site identification and patient medical condition will be used.

3.12 Reporting

3.12.1 Patient Disposition

The number and percentage of patients will be tabulated for each of the following categories by disease group:

- Screened (total only);
- Full Analysis Set (mITT population); overall and by disease group
- PPS population; overall and by disease group
- Safety population;
- Patients who complete the study;
- Patients who has early termination of protocol and reason for early termination
- Patients who were successfully screened and enrolled but did not start the study protocol (excluded from the FAS).
- Protocol deviations will be summarized by deviation type

3.12.2 Baseline Demographics

Baseline demographics will be presented for the overall population as well as stratified by disease group and by site. Summary measures will include percentage for categorical variables and lower 25%, median, upper 75%, mean and standard deviation for continuous variables.

Along with the demographic data, the following disease characteristics will be summarized for the FAS:

- Inpatient (Acute vs. inpatient rehabilitation)/Outpatient split
- Inpatient primary admission diagnosis
- Inpatient secondary admission diagnosis
- Invasive mechanically ventilated patients
- Outpatient reason for referral for VFSS
- Outpatient previous diagnoses
- NIH Stroke scale (stroke subjects only)

Baseline and all above patient characteristics will be presented for the FAS overall population as well as stratified by primary daignosis group (Stroke, Other Neuro and Other).

3.12.3 Study Administration

For each consistency and swallow count the number of subjects completing that particular swallow will be reported. This will be stratified by consistency and disease group (also overall).

3.12.4 Efficacy Analysis

For each of impaired safety and efficiecy, point estimates and confidence interval will be presented for sensitivities and specificities by bolus consistencies. This will be presented for the overall FAS. P-value for testing the efficacy hypothesis in Equation (1) will be reported for the primary efficacy endpoint. In addition, if the primary efficacy hypothesis test meets statistical significance then p-values for secondary efficacy analysis will also be presented in a hierarchical fashion until the first non-significant secondary hypothesis test. The order of the hierarchy is given in Section 2.4.2. The significance level for efficacy analysis at each interim and final analysis will be obtained using the LDOF spending function [2].

3.12.5 Prevalence of Swallow Safety and Efficiency Impairments

Bolus level prevalence of Safety and Efficacy impairments will be summarized by swallow consistency and by swallow count. The number of missing VFSS label will also be summarized.

Subject level prevalence will also be reported similarly by swallow consistency and swallow count. Here, only the first impairment will be counted. Thus the prevalence at a particular swallow count is defined as the number of subjects showing their first impairment at that swallow count divided by the number of subjects at risk of impairment at that swallow. Cumulative prevalence will also be summarized.

Prevalence tables will also be reported stratified by primary diagnosis group and by site (in all 13 sites).

3.12.6 Safety Analysis

All safety analysis will be performed on the Safety Analysis Set. All adverse Events (AEs) will be coded using the Medical Dicotionary for Regulatory Agencies (MedDRA). A TEAE is defined as an AE that starts or worsens on or after study Day 1 (defined as the procedure day), and no more than XX days after the procedure. All TEAEs will be listed by subject number and MedDRA coding.

The number and percentage of subjects with TEAE will be summarized in several different tables:

- All TEATs by System Organ Class (SOC) and preferred term (PT)
- Study procedure related AEs by SOC and PT
- Severity of All TEAEs by SOC and PT.

A subject with 2 or more AEs within the same level of the MedDRA term will be counted only once in that level using the most extreme incident (most severe for the severity table and related for the relationship to study procedure table).

4 Blinding and Firewall for Interim Analysis

The key blinding required for this trial is the blinding of the sites to the VFSS results determined by a central VFSS laboratory and the blinding of the central VFSS laboratory to the yet-to-be validated results from the DDS available at the sites. There will be no communication between the sites and the central laboratory. The results of the VFSS on each subjects' boluses will be kept blinded and would not be available to the sites or the sponsors. It will be made available to the iDMC and the iSC members with the unblinded role through a secured firewall system immediately prior to the time of the interim and the final analyses. The details will be pre-specified in the study protocol as well as in the DMC charter.

Cytel's Access Control Execution System ACES will be used as the firewall system. This system stores all study conduct and analysis related documents in a secured and traceable way. All members of the sponsor team, iDMC, iSC and others as required are granted accounts on this system with a specified role. The roles of iDMC and iSC members carrying out the unblinded analysis at the interims will be given the "unblinded" role. All other account holders on ACES are given a "blinded" role. This is administered by the ACES administrator who also has an "unblinded" role. Members also may or may not have the rights to upload documents on the ACES system for that particular trial. Any document uploaded by a member also has a specified role "unblinded" or "blinded". The correctness of the this role assigned to the document is checked automatically as well as manually by the ACES administrator before being available according to the role to other ACES accounts. Naturally, only a user with an "unblinded" role can upload a document with an "unblinded" role.

Prior to the interim analysis iDMC meeting, all data except the VFSS data that is maintained by the database manager will be uploaded to the ACES and analysis datasets will be programmed in a blinded fashion using dummy VFSS results. Once the interim analysis is ready to be carried out, the iDMC will be required to send an notification to the central VFSS laboratory to upload the VFSS results on ACES. This document will have a "unblinded" role and will only be available to the iDMC members and the iSC members

engaged in producing the interim analysis results. Once this document is uploaded in the ACES system there will be no direct communication between the sponsors and/or the sites and the "unblinded" members. All communication will then need to occur within the ACES system. The data remains blinded to the sponsors and the site until the end of the final analysis (planned or in case of an early stop).

With ACES it is possible to generate a report on all documents in the ACES system which minimally includes the following information:

1. Time of upload
2. Role specified
3. List of all users who have seen or had access to each particular document with the corresponding time

The general ACES methodology and process is summarized in the Figure 4-1.

References

[1] C. R. and Pocock Mehta S. J., *Adaptive increase in sample size when interim results are promising: a practical guide with examples*, Stat. Med. **30** (2011), no. 28, 3267–3284.

[2] D. L. and Lan DeMets K. K., *Interim analysis: the alpha spending function approach*, Stat. Med. **13** (1994), no. 13-14, 1341–1352.

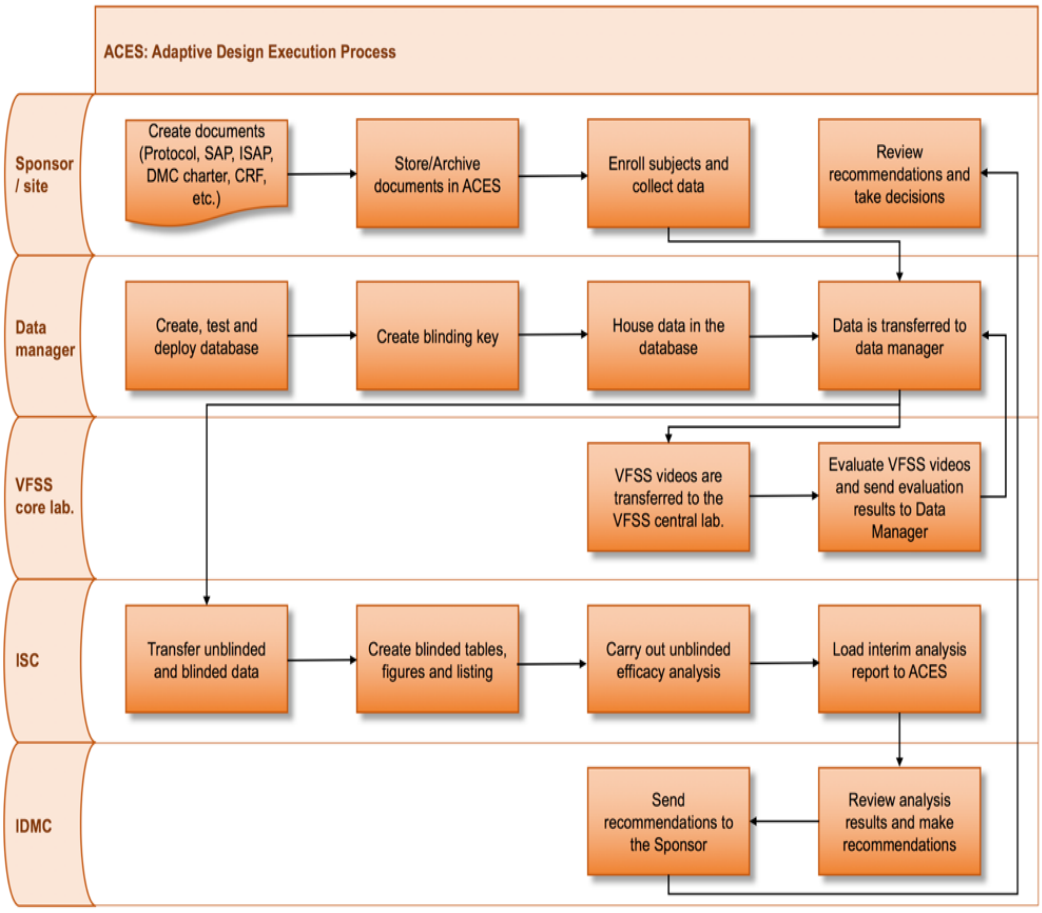

Figure 4-1: The ACES System to be followed for the PORSCHE Trial

5 Appendix

5.1 Simulations with Sample Size Re-estimation

Simulations for the proposed GSD with SSR in Section 5 were carried out under the null hypothesis (sensitivity $\leq .8$ and/or specificity $\leq .5$) while varying the prevalence from 35% to 65%. Power from 10,000 monte-carlo runs for different combinations of sensitivity and specificity in Table 8-1 show that the type-I error is preserved at the 2.5%.

Simulations under the alternative (sensitivity $> .8$ and specificity $> .5$ for the primary and the first two secondary endpoints) show that for prevalence varying between 35% and 65%, sensitivity around 90% and specificity greater than 60% the power achieved is greater than 90% for a expected sample size between 950 and 1050 subjects under the adaptive group sequential design.

Table 5-1: Simulation results under the Union Null Hypothesis for proposed group sequential design with thresholding at interim-1 and SSR at the last interim, for varying prevalence, sensitivity and specificity.

Prev.	Se	Sp	Prop. ChangeTH	Prop. Futility stop at IA1	Overall Power	Min.N	Avg.N	Max.N
0.35	0.80	0.50	0.05	0.93	0.0004	876	882	1300
0.50	0.80	0.50	0.05	0.94	0.0001	876	882	1300
0.65	0.80	0.50	0.07	0.92	0.0000	876	884	1300
0.35	0.85	0.50	0.19	0.77	0.0094	876	904	1300
0.50	0.85	0.50	0.18	0.78	0.0106	876	904	1300
0.65	0.85	0.50	0.21	0.75	0.0137	876	909	1300
0.35	0.90	0.49	0.40	0.57	0.0150	876	930	1300
0.50	0.90	0.49	0.39	0.58	0.0160	876	929	1300
0.65	0.90	0.49	0.40	0.57	0.0144	876	931	1300
0.35	0.80	0.55	0.16	0.78	0.0088	876	900	1300
0.50	0.80	0.55	0.16	0.80	0.0067	876	899	1300
0.65	0.80	0.55	0.19	0.77	0.0046	876	902	1300
0.35	0.79	0.60	0.33	0.63	0.0062	876	916	1300
0.50	0.79	0.60	0.35	0.61	0.0052	876	919	1300
0.65	0.79	0.60	0.35	0.62	0.0037	876	918	1300

Prev.: Data generating prevalence
Se: Data generating Sensitivity
Sp: Data generating Specificity
Prop ChangeTH: Proportion of runs where threshold was changed at IA-1
Prop. Futility Stop at IA-1: Futility Futility stopping triggered at IA-1, however simulations carried out under non-binding futility
Overall Power: Power to detect at least the primary hypothesis
Min.N, Avg.N, Max.N: All calculated over 10,000 runs with non-binding futility

Table 5-2: Simulation results under the Intersection Alternate Hypothesis for proposed group sequential design with thresholding at interim-1 and SSR at the last interim, for varying prevalence, sensitivity and specificity.

Prev.	Se	Sp	Prop. ChangeTH	Overall Power	Power to Reject All 3 endpoints	Min.N	Avg.N	Max.N
0.35	0.90	0.60	0.34	0.995	0.894	250	969	1300
0.45	0.90	0.60	0.33	0.993	0.903	250	946	1300
0.55	0.90	0.60	0.33	0.979	0.901	250	989	1300
0.65	0.90	0.60	0.36	0.947	0.829	250	1044	1300
0.35	0.85	0.65	0.74	0.625	0.258	525	1091	1300
0.45	0.85	0.65	0.65	0.741	0.406	525	1100	1300
0.55	0.85	0.65	0.57	0.829	0.538	525	1105	1300
0.65	0.85	0.65	0.48	0.881	0.614	525	1103	1300
0.35	0.92	0.55	0.56	0.705	0.342	525	1082	1300
0.45	0.92	0.55	0.62	0.647	0.275	525	1079	1300
0.55	0.92	0.55	0.64	0.580	0.192	525	1070	1300
0.65	0.92	0.55	0.63	0.481	0.098	876	1054	1300
0.35	0.92	0.58	0.41	0.969	0.828	525	1026	1300
0.45	0.92	0.58	0.46	0.941	0.803	250	1050	1300
0.55	0.92	0.58	0.50	0.908	0.734	525	1080	1300
0.65	0.92	0.58	0.54	0.836	0.580	525	1089	1300

Prev.: Data generating prevalence
Se: Data generating Sensitivity
Sp: Data generating Specificity
Prop ChangeTH: Proportion of runs where threshold was changed at IA-1
Overall Power: Power to detect at least the primary hypothesis
Min.N, Avg.N, Max.N: All calculated over 10,000

5.2 Interim Analysis Plan Flowchart

The different interim analysis plans are summarized in Figure 5.1 below.

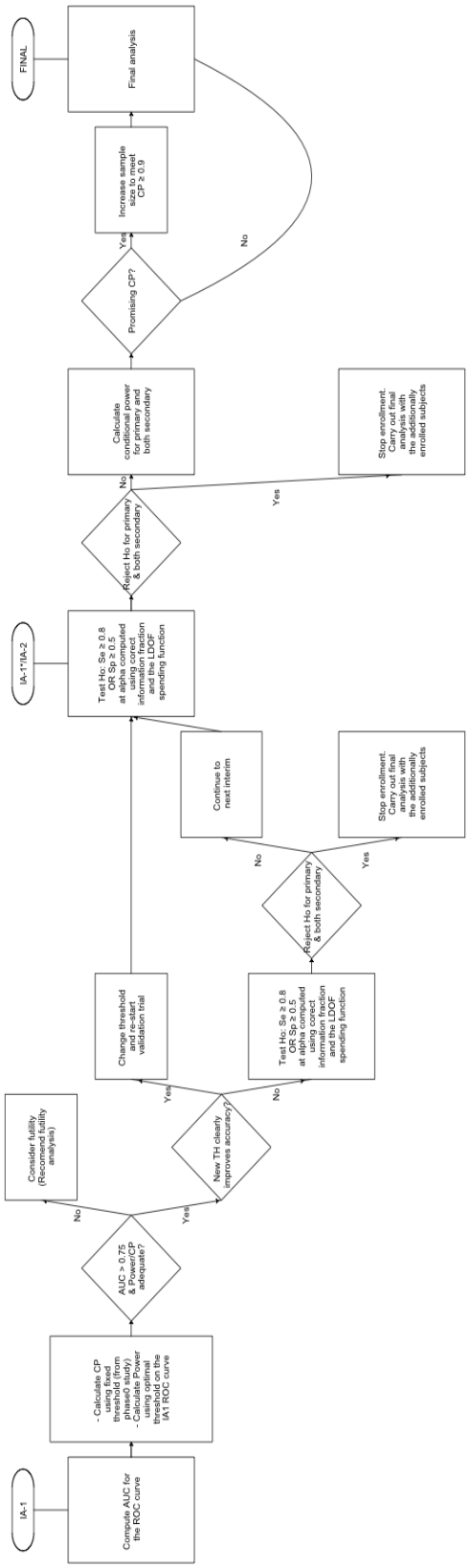

Figure 5-1: Flowchart for Interin Analysis Plans