

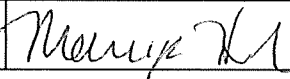
Tornier, Inc.

Pyrocarbon IDE Study

IDE No. 15A-T-PYC-R

Statistical Analysis Plan

Author:

Title	Name	Signature	Date
Principal Statistician, Techonomics Research	Manya Harsch		10 Jan 2019

Approvers:

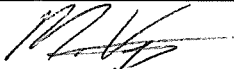
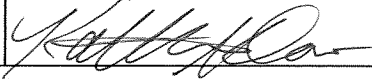
Title	Name	Signature	Date
Clinical Study Manager, Wright	Matt Venegoni		11 Jan 2019
Clinical Research Manager, Wright	Kathleen Davis		11 Jan 2019

Table of Contents

1.	Purpose	3
2.	Scope.....	3
3.	Applicable Documents	3
4.	Software.....	3
5.	Definitions/Acronyms	3
6.	Trial Objectives.....	5
7.	Trial Hypotheses	5
8.	Trial Success Criteria.....	5
9.	Trial Design.....	5
9.1.	Randomization	6
9.2.	Blinding.....	6
10.	Sample Size Considerations	7
11.	Data Structure and Handling	7
11.1.	Data Handling and Transfer	7
11.2.	Missing Data and Censoring	7
11.3.	Visit Windows	8
12.	Statistical Analyses.....	8
12.1.	General Considerations	8
12.2.	Analysis Populations.....	8
12.3.	Subject Disposition.....	9
12.4.	Demographics and Baseline Characteristics	9
12.5.	Primary Effectiveness Analysis	9
12.5.1.	Primary analysis.....	9
12.5.2.	Sensitivity Analysis	10
12.5.3.	Pooling Data Analyses	10
12.6.	Inferential Secondary Endpoints.....	10
12.6.1.	Constant-Murley Score	12
12.6.2.	ASES Score.....	13
12.6.3.	EQ-5D Scoring	13
12.7.	Additional Secondary Endpoints	14
12.8.	Exploratory Analyses	14
12.9.	Sex Analysis.....	14
12.10.	Gap Measurement Analysis	14
12.11.	Safety Analysis	15
12.12.	Other Data	15
13.	Comparability to Aequalis Dataset	15
13.1.	Aequalis Data	16
13.2.	Propensity Score Analysis	16
14.	Version History.....	18
15.	References.....	18

1. Purpose

This statistical analysis plan (SAP) describes the statistical methods to be used during the reporting and analysis of data collected under the Tornier, Inc., Pyrocarbon IDE Study protocol.

2. Scope

This SAP should be read in conjunction with the study protocol and electronic case report forms (eCRF). This version of the plan has been developed with respect to the G140202 protocol version 5, dated 10SEP2015. Any further changes to the protocol or eCRF may necessitate updates to the SAP.

3. Applicable Documents

Document Number	Document Title
IDE No. G140202, v 6	Pyrocarbon IDE Study, version 6 (Protocol #15A-T-PYC-R)
	eCRFs - as of 06 January 2017

4. Software

All tables, listings, and figures will be produced using SAS Version 9.3 (SAS Institute, Cary, NC.) or a later version of SAS. In the event an analysis is required, that is better suited for a statistical package other than SAS, this other package (e.g., R) will be used.

5. Definitions/Acronyms

Acronym	Definition
ADL	Activities of Daily Living
ANOVA	Analysis of variance
AE	Adverse Event
ASES	American Shoulder and Elbow Surgeons Standardized Shoulder Assessment
CIP	Clinical Investigational Plan
CRO	Clinical Research Organization
CRF	Case Report Form
DOS	Day of Surgery
eCRFs	Electronic Case Report Forms
EDC	Electronic Data Capture
EQ5D	EQ-5D™ is a standardized instrument for use as a measure of health outcome.
FCI	Functional Comorbidity Index
FDA	Food and Drug Administration
GCP	Good Clinical Practice
HA	Hemiarthroplasty
HH	Humeral Head
ICF	Informed Consent Form
ITT	Intent-to-treat
OUS	Outside of the United States of America
PI	Principal Investigator
PP	Per protocol
PyC	Pyrocarbon

QOL	Quality of Life
ROM	Range of Motion
SAE	Serious Adverse Event or birth defect.
SANE	Single Assessment Numeric Evaluation Score
SST	Simple Shoulder Test
TSA	Total Shoulder Arthroplasty
UADE	Unanticipated Adverse Device Effect
US	United States of America
VAS	Visual Analog Scale

6. Trial Objectives

The purpose of this clinical study is to evaluate the safety and efficacy of the Aequalis™ Pyrocarbon Humeral Head (Pyrocarbon HH). The data generated by this study are intended to provide adequate safety and effectiveness information necessary to support a Food and Drug Administration (FDA) submission for device clearance. Data from this clinical study may be used to support future regulatory submissions, including those outside of the United States (US).

The study is designed to evaluate the safety and efficacy of the Pyrocarbon HH when used with the FDA cleared Ascend Flex System in the primary replacement of the humeral side of the shoulder joint.

7. Trial Hypotheses

The objective of the primary endpoint analysis is to demonstrate that the rate of the composite success endpoint in subjects with the Pyrocarbon HH is non-inferior to the rate observed in the Tornier Aequalis Post-Market Outcomes Study dataset. Formally, the hypothesis to be tested is:

$H_0: p \leq PG - \delta$

$H_A: p > PG - \delta$

Where p is the proportion of subjects with a success for the composite success endpoint, PG is the performance goal, and δ is the non-inferiority margin. The PG is derived from data for the Aequalis subjects with the appropriate indications and with two years of follow up data (or revision prior to two years). The PG and δ being used for the hypothesis test take on the values of 0.85 (or 85%) and 0.10 (or 10%), respectively.

8. Trial Success Criteria

The non-inferiority of Pyrocarbon HH will be evaluated by a two-sided 95% confidence interval for the success rate (p). The study will be considered a success if the lower bound of the confidence interval is greater than 0.75 (0.85 – 0.10).

9. Trial Design

This is a multi-center, prospective, single arm, investigational clinical study. The study will be conducted at up to 20 sites in the US. Each participating site will be encouraged to attempt implants in approximately 5 subjects in a 12 month time period, or until the enrollment limit is met. A subject is “enrolled” and count towards the enrollment limit of 190 subjects once the consent form is signed. A subject will count towards the limit of 157 subjects when inclusion/exclusion criteria are met (including unsuccessful and successful implantation of a PYC HH). Up to 157 subjects will have attempted implants with a Pyrocarbon HH to ensure at least 133 evaluable subjects are available for the primary endpoint analysis.

It is anticipated that this study will require approximately up to 24 months for subject enrollment. To ensure data are adequately distributed among sites and geographies, no more than 20% ($n=31$) of subjects may occur at a single site. Each site should attempt to implant 5 subjects within 12 months of activation. There is no minimum requirement for enrollments per site.

All enrolled subjects will be followed for at least 24 months (except for pre-operative screen failures). Study duration (first site activation to final subject follow-up) is expected to be approximately 4 years. If deemed necessary, subjects may be asked to return for additional visits beyond their final protocol visit to collect long-term data

Subjects who are enrolled and have a successful implant attempt (implant stable through wound closure) will have study assessments at 2 weeks, 3 months, 6 months, 12 months, 24 months, and annually thereafter (until the last implant attempt subject completes their 24 month visit) following surgery. Subjects will exit the study after the last implant attempt subjects completes their 24 month assessment unless it is determined, at the end of the study, to follow subjects long-term.

Subjects who are enrolled and have an unsuccessful implant attempt (attempt to implant the Pyrocarbon Humeral Head was not successful during the index procedure) will have annual post-op visits at 12 and 24 months in order to collect any available safety information. This also includes subjects who are excluded due to excessive glenoid reaming performed with the intent of implanting the Pyrocarbon Humeral Head. Subjects will exit the study after the 24 month assessments are complete.

Subjects who are enrolled but become screen failures intra-operatively (prior to the implant attempt) will have the reason for screen failure documented and then will be followed annually to collect any available safety information. Subjects will exit the study after the 24 month assessments are complete.

Subjects who are enrolled but become screen failures pre-operatively (prior to surgical procedure) will have the reason for screen failure documented then they will be exited from the study.

If any component of the investigational system is revised, the subject will follow the planned follow-up visits. If any component of the investigational system is permanently explanted during the course of the study, the subject will only have annual post-op visits at 12 and 24 month to gather available safety information. Revisions and explants are defined in section 11.2.

Data collection schedule is as follows:

	Baseline (Pre-op)	Day of Surgery (DOS)	Post-op Follow-up (5-21 days)	Follow-up 3M (69-111 days)	Follow-up 6M (159-201 days)	Follow-up 12M (323-407 days)	Follow-up 24M (688-772 days)	Follow-up Annual (if needed)
Demographics, Medical History and Indication	X							
Intraoperative Data		X						
Subject Questionnaire	X			X	X	X	X	X
Range of Motion	X			X*	X*	X	X	X
Strength Test	X			X*	X*	X	X	X
Pain Medications	X	X	X	X	X	X	X	X
Adverse Event Assessment		X	X	X	X	X	X	X
X-ray	X		X		X	X	X	X

*Optional

9.1. Randomization

There is no randomization as this is a single arm study.

9.2. Blinding

There is no blinding as this is a single arm study.

10. Sample Size Considerations

Statistical considerations for powering the primary endpoint include the following:

- 80% power with a one-sided 0.025 level of significance
- Non-inferiority test of one proportion
- 10% non-inferiority margin, and performance goal of 85%
- Assumed investigational system success rate of 85%
- Attrition rate of 15%

The sample size for this endpoint was calculated using SAS (Version 9.3) under a one-sided z-test of a binomial proportion. The minimum required sample size is estimated using the above list of statistical considerations. The resultant sample size under these assumptions is 133 subjects. With 15% attrition, 157 total subjects with an attempted implant are needed. Up to 190 subjects may be enrolled to account for pre and intra-operative screen failures. Enrollment will stop once 157 subjects have had an implant attempt.

11. Data Structure and Handling

11.1. *Data Handling and Transfer*

Data management will be undertaken by Tornier. A biostatistician from an independent CRO will either be provided access to download SAS datasets or Tornier will provide them upon request.

Programming of analysis datasets, tables, figures and listings will be conducted during the data management phase of the study. Tables, figures, and listings may be reviewed prior to final data lock for data review. Any data values requiring investigation or correction will be identified, and protocol deviations will be reviewed. The final run of outputs will take place after the data is deemed final. The study team will not have access to the analysis data sets until the final data set is available.

11.2. *Missing Data and Censoring*

The impact of missing data for the primary efficacy endpoint will be assessed by a sensitivity analysis. See the primary endpoint section for details.

No imputation for missing data will be performed on secondary endpoint data; i.e., patients missing the endpoint result at a visit or prematurely withdrawing prior to the time point at which the endpoint is measured will be excluded from the analysis of that endpoint.

Revisions:

A subject who has undergone a study system revision will be considered a failure for the primary efficacy endpoint. All available safety data will be included in summaries.

A revision is a procedure that adjusts or in any way modifies or removes part of the original implant configuration, with or without replacement of a component, after the initial surgery. A revision may also include adjusting the position of the original configuration. This may include removing a component of a joint implant and only replacing that single component. If a subject has revision surgery, the subject shall remain in the study and complete all required follow-up visits.

If a subject has a revision that includes permanent removal of the Pyrocarbon HH, they shall remain in the study and complete all required annual follow-up visits.

Outcome Questionnaires:

In order to include patients in the analyses who were missing a small subset of responses to an outcome questionnaire, rules were implemented to maximize the number of patients with scores calculated. Missing responses are effectively imputed as 0. Specific rules for the Constant, ASES and EQ5D are located in the sections pertaining to the respective score later in this document.

11.3. Visit Windows

All data attributed to a time point per the CRF will be included in the analysis of that time point, regardless of whether it is out of the protocol specified window. A summary of visit-window adherence will be included in the final analysis. Unscheduled visit data will be included in adverse event and other summaries that are not specific to a time point. The table below outlines the visit windows for protocol prescribed visits.

Study Visit	Visit Window
Baseline	60 days pre-op through DOS
Post-op Follow-up	5-21 days post-op
3 Month Follow-up	90 Days post-op \pm 21 days
6 Month Follow-up	180 Days post-op \pm 21 days
12 Month Follow-up	365 Days post-op \pm 42 days
24 Month Follow-up	730 Days post-op \pm 42 days
36 Month Follow-up (if needed)	1095 Days post-op \pm 84 days
48 Month Follow-up (if needed)	1460 Days post-op \pm 84 days
60 Month Follow-up (if needed)	1825 Days post-op \pm 84 days
72 Month Follow-up (if needed)	2190 Days post-op \pm 84 days

12. Statistical Analyses

12.1. General Considerations

Data will be summarized using descriptive statistics. Continuous data will be summarized using mean, standard deviation, minimum, and maximum. Where appropriate, non-parametric measures of location (e.g. median, interquartile range) may be provided. For categorical data frequencies and percentages will be provided.

The surgery date will be considered study day 1.

12.2. Analysis Populations

The Intent-to-Treat (ITT) population will consist of all subjects who have undergone an attempted or successful implantation of the investigational system.

The Per-Protocol (PP) population will include all subjects who have undergone successful implantation of the investigational system, completed 24 months of follow-up, and have had no major protocol violations (defined as not meeting inclusion/exclusion criteria or not being consented properly). Major protocol violations (inclusion/exclusion criteria not met or consent issues) will be determined by the inclusion/exclusion and protocol deviation data sets.

Effectiveness analyses will be performed on the ITT population. Analysis of the PP population will be used in support of the ITT analyses of the primary efficacy endpoint. The safety analysis will be performed on the ITT population.

12.3. *Subject Disposition*

Subject disposition will be presented by:

- Summary of patient accountability following the FDA guidance Clinical Data Presentations for Orthopedic Device Applications (2004).
- Summary of early withdrawal and reason for early withdrawal.

12.4. *Demographics and Baseline Characteristics*

Demographics and baseline characteristics will be assessed by descriptive statistics. These factors will include (but not be limited to):

- Age
- Gender
- Body Mass Index (BMI)
- Medical history (including diagnosis)
- Adjusted Constant score
- Pain

12.5. *Primary Effectiveness Analysis*

12.5.1. *Primary analysis*

The primary endpoint is the rate of patient success at 24 months. A subject is a success at 24 months if:

- Their change in Constant score is ≥ 17 and
- They did not have revision surgery; and
- There is no radiographic evidence of system disassembly or fracture, and
- They did not have a system-related serious adverse event.

Subjects that undergo revision surgery within 24 months will be considered a primary endpoint failure at the time any component of the system is revised.

If a subject is missing the 24 month radiographic evidence assessment or Constant score component of the composite endpoint and did not have a system-related SAE or revision surgery, they will be excluded from the primary analysis due to inability to claim success or failure.

Subjects in whom an implant was attempted but the implant was unsuccessful will be included as failures in the primary analysis, regardless of the availability of 24 month radiographic evidence assessment or Constant score.

The presence of radiographic evidence of system disassembly or fracture will be assessed by independent radiologists. Discrepancies among the independent evaluations will be settled by majority rule. If the radiologists are unable to assess or the image wasn't available at 24 months, this will be considered missing data and the subject will be excluded from the primary analysis.

The non-inferiority of the test system to the performance goal will be evaluated using the 95% CI for the observed test system success rate. The study will be considered a success if the lower bound of the confidence interval is greater than 75% (85% - 10%).

The number of subjects who are successes (X) for each of the 4 individual components of the composite success endpoint, along with the number evaluated (N), will be summarized at 12 and 24 months as (X/N) and displayed with the overall success endpoint results based on data up to 12 and 24 months.

12.5.2. Sensitivity Analysis

Two separate sensitivity analyses will be completed on the primary endpoint, outlined below:

- 1) The primary effectiveness analysis will be repeated under the assumption that all subjects with missing endpoint data were successful.
- 2) The primary effectiveness analysis will be repeated under the assumption that all subjects with missing endpoint data were failures.
- 3) The primary effectiveness analysis will be repeated using multiple imputation for missing endpoint data. The covariates that may (list not all-inclusive) be included in the multiple imputation process are: age, gender, treated shoulder, arthritis diagnosis, Constant score (baseline and 12 month), and BMI.

12.5.3. Pooling Data Analyses

A pooling analysis will be performed to assess the poolability of sites with respect to the primary endpoint measure. Success rates will be summarized by site, according to the primary analysis method. A Fisher's Exact Test will be used to test for a difference in composite success rate across sites. A significance level of 0.1 will be used to determine whether the sites are poolable. If a significant difference is found between sites, additional analyses will be done to identify factors that may be associated with this difference. Additionally, if a significant difference is observed, a random site adjusted estimate of the composite success rate, along with 95% CI, will be provided.

Although enrollment will be monitored in an effort to strive for even allocation between sites, those sites with less than five (5) evaluable subjects will be combined into a pseudo-site for purposes of analysis. To protect against having an overly large pseudo-site, when one pseudo-site exceeds five (5) evaluable subjects, a second pseudo-site will be formed. This process will continue as needed each time a pseudo-site exceeds five (5) evaluable subjects.

12.6. Inferential Secondary Endpoints

The following secondary data will be evaluated for significant change from baseline to 24 months. Analyses will be performed on the ITT population, using all available data. The overall type I error will be controlled using the Hochberg method for adjusting for multiple comparisons, and will be tested only if the primary endpoint is met. Endpoints will be evaluated at a one-sided alpha level of 0.025.

- Constant score and Adjusted Constant score (see Section 12.6.1)
- American Shoulder and Elbow Surgeons (ASES) Score (see Section 12.6.2)
- Single Assessment Numeric Evaluation (SANE)
 - The SANE rating is determined by the by the subject's written response to the following question "How would you rate your shoulder today as a percentage of normal (0% to 100% scale with 100% being normal)?"

- EQ-5D (see Section 12.6.3)
- Pain (VAS)
 - Based on 0-10 scale
- Range of Motion
 - Forward Flexion
 - Abduction
 - Internal rotation
 - External rotation
- Strength (average pull strength)
- Revision Rate

The hypothesis test for each of the above endpoints, with the exception of Pain (VAS) and revision rate, will be of the form:

$$H_0: \mu \leq 0$$

$$H_A: \mu > 0$$

Where μ is the mean change from baseline to 24 months (follow-up minus baseline) for that endpoint. The hypothesis test for the Pain (VAS) endpoint will be of the form:

$$H_0: \mu \geq 0$$

$$H_A: \mu < 0$$

Where μ is the mean change from baseline to 24 months (follow-up minus baseline) in the VAS Pain score.

For each endpoint excluding revision rate, descriptive statistics will be provided for subjects at baseline and 24 month follow-up. The difference from baseline to 24 months will be summarized using descriptive statistics, and a paired t-test will be performed. Additionally, the 95% confidence interval will be calculated for the change from baseline.

The hypothesis test for the Revision rate will be of the form:

$$H_0: p \geq 13.2\%$$

$$H_A: p < 13.2\%$$

Where p is the proportion of subjects that had a revision in the 24 month time period. This endpoint was added based upon FDA feedback. As the sample size is already established at 133 subjects, the PG is based upon the observed revision rate from the Aequalis dataset and a non-inferiority margin of 8% that has approximately 90% power with a 1-sided alpha of 0.025 with the sample size of 133.

In the Aequalis historical control dataset, 10 of 191 subjects had a revision of some type within 24 months from surgery for an estimated revision rate of 5.2%. Using the same non-inferiority margin of 10%, a sample size of 100 is adequate to test with at least 90% power and a one-sided alpha of 0.025. Estimates for the revision rate at 24 months and an upper 97.5% confidence bound will be generated using a Kaplan-Meier time to event analysis. If this value is less than 13.2%, then the null hypothesis is rejected and it is concluded that the revision rate is acceptable and not inferior to prior approved devices. Additional summaries

will include the proportion of subjects with a revision rate within 24 months along with a binomial two-sided 95% confidence limits.

12.6.1. Constant-Murley Score

The Constant (Constant 2008) and adjusted Constant score (Constant 1986) will be summarized as part of the primary endpoint and as an inferential secondary endpoints. The Constant score is calculated as the sum of the following components:

- Average of 2 pain scores (15 points max):
 - Pain in shoulder normal activity: No=15, mild=10, moderate=5, severe=0
 - Linear 0-15 scale, where 0=no pain and 15=max pain, scored as the inverse of the scale: 15-(scale score); e.g., 15-0=15 (no pain contributes 15), 15-1=14,...,15-15=0 (max pain contributes 0).
- Sum of 4 activities of daily living questions (20 points max):
 - Occupation or daily living limited by shoulder: no=4, moderate=2, severe=0
 - Leisure and recreational activities limited by shoulder: no=4, moderate=2, severe=0
 - Sleep disturbed by shoulder: no=2, sometimes=1, yes=0
 - Highest level can use arm for reasonably painless activity: waist=2, xiphoid=4, neck=6, top of head=8, above head=10
- Sum of 4 ROM measure (40 points max):
 - Forward flexion: 0-30=0, 31-60=2, 61-90=4, 91-120=6, 121-150=8, >150=10
 - Abduction: 0-30=0, 31-60=2, 61-90=4, 91-120=6, 121-150=8, >150=10
 - External rotation: 2 points for each location reached (up to 10 points total)
 - Internal rotation: lateral thigh=0, buttocks=2, lumbrosacral junction=4, waist=6, 12th dorsal vertebra=8, interscapular region=10
- Power score (25 points max):
 - Average of 3 pulls (maximum of 25 pounds per pull) [Note: Statistician to verify pull was performed since some may say no but have 0 recorded.]

The average of three pulls were planned to be used to determine the strength component of the Constant score. In order to minimize missing data, the score will be calculated as long as at least one pull recorded. The average of the available pulls (1-3) will be used in the calculation.

Any combination of missing values that equate to a potential of >20 points would be excluded. For example, missing strength testing excludes a patient (if the subject was unable to complete the strength test then this would not exclude the patient); missing more than two of the ROM measures would exclude a patient; missing pain and one ROM test would exclude a patient, etc. For questionnaires missing partial data but not meeting the

exclusion threshold, missing responses are assumed to be worst case (contribute 0). Point details are as follows:

- Pain (up to 15 points): only one of the two pain scores is required
- Activities of daily living: 4 questions; total of 20 points
- ROM: 4 tests; 10 points each
- Strength: Strength contributes 25% of the total score (>20% threshold).

The constant score adjustments will be made using Constant adjustment values in Table 1:

Table 1: Constant Score Adjustments

Age (y)	Male	Female
21-30	98	97
31-40	93	90
41-50	92	80
51-60	90	73
61-70	83	70
71-80	75	69
81-90	66	64
91-100	56	52

Apply the adjustment by dividing the score by the factor in the table, then multiplying by 100.

12.6.2. ASES Score

The maximum shoulder score is 100 points and is derived by the following formula (Michener 2000):

$$[(10 - VAS \text{ score for pain}) \times 5] + [(5/3) \times \text{Cumulative ADL score}]$$

in which the VAS score ranges from 0 to 10 (cm) and the 10 ADL questions are scored as: 0='unable to do', 1='very difficult to do', 2='somewhat difficult to do', 3='not difficult'. To account for partially missing questionnaires, the following rules will be applied:

- Pain question required. Rationale: it is worth up to 50 points of the 100 point calculation.
- Up to 4 of the 10 patient questionnaire questions can be missing; if 4 or fewer are missing the missing response(s) is assumed to be worst case (contribute 0). Rationale: 4 patient questionnaire answers contribute 20% of the total score.

12.6.3. EQ-5D Scoring

The EQ-5D will be summarized using EQ-5D-5L descriptive score and EQ VAS score (<http://www.euroqol.org/>).

The EQ-5D-5L score will be derived by recoding the 5 dimensions to a single-digit numeric value based on the level they indicated for that dimension. These digits will be concatenated (in the order of the actual dimensions) together to obtain a 5-digit number describing the subjects state of health. This number will be used to allocate a score for each subject using a downloaded spreadsheet from aforementioned website. Subject scores will be determined from the spreadsheet based on their 5-digit number. Subjects must have all 5 dimensions in order to obtain an EQ-5D-5L score.

The EQ VAS score is derived from the respondent's self-rated health on a 20 cm vertical, visual analogue scale with endpoints labelled 'the best health you can imagine' and 'the worst health you can imagine'. This information is used as a quantitative measure of health as judged by the individual respondents.

12.7. *Additional Secondary Endpoints*

The following secondary endpoints will be presented using descriptive statistics only. Analyses will be performed on the ITT population, using all available data.

- Adverse events
- Revision rate
- Level of satisfaction with the shoulder
- X-ray data: glenohumeral joint space width, glenoid osteophytes, glenoid morphology, glenoid erosion, and glenoid wear, humeral component radiolucency, osteolysis, migration, subsidence, subluxation, humeral head integrity, acromial humeral distance, anatomic fracture, and additional observations

The revision rate and radiographic data will be summarized and qualitatively compared to a subset of data from the Tornier Aequalis Post-Market Outcomes Study and data reported in literature. This will be included in the final clinical report.

Note that safety is included in the primary composite endpoint and also as a descriptive secondary endpoint. See section 12.10 below for additional details on planned safety analyses.

12.8. *Exploratory Analyses*

Additional, ad hoc exploratory analyses may also be conducted.

12.9. *Sex Analysis*

Potential differences between genders in the primary endpoint success rate will be evaluated. The primary endpoint success rate will be calculated and presented for each sex. A comparison between the sex will be made using Fisher's exact test.

12.10. *Gap Measurement Analysis*

Gap measurements will be collected on all devices. During assembly of the Pyrocarbon humeral head onto the male taper a "gap" measurement is used to determine the acceptability of component tolerancing prior to assembly. The gap measurements will be summarized using descriptive statistics, including median and interquartile range. A subgroup analysis will be performed where the primary endpoint will be summarized by quartiles of the gap measurements. A comparison between the subgroups will be made using Fisher's exact test.

12.11. Safety Analysis

All adverse events and deaths will be reported and summarized. This includes all operative procedure and postoperative complications. Events are categorized as device-related, procedure-related) and systemic (non-device or procedure related). An event is considered related if the investigator indicated it was definitely or possibly related.

Safety analyses will be performed on the ITT population. The objective of the safety analysis is to demonstrate that the investigational system has an acceptable safety profile. Adverse events will be summarized by the proportion of subjects with serious adverse events, device-related adverse events and unanticipated adverse device effects (from adjudication), and also by the individual AE terms. Additionally, a summary will be presented according to the FDA guidance Clinical Data Presentations for Orthopedic Device Applications (2004), separating events by category (device-related, procedure-related, systemic) and by time period. The denominator for the time periods will include all subjects who were evaluated at some point during the time period (scheduled visit or AE within window). The number of occurrences of each type of adverse event within the visit window will be presented. For this analysis, visits are defined as follows, using the upper end of the planned visit windows from the protocol:

- Operation: day of procedure (Day 1)
- Post-op: Day 2 to Day 21
- 3 months (12 weeks): Day 22 – Day 105
- 6 months (26 weeks): Day 106 – Day 203
- 12 months (52 weeks): Day 204 – Day 392
- 24 months (110 weeks): \geq Day 393
- A total column may also be added

Subsequent secondary surgical interventions will be presented similarly to the above adverse event description following the FDA guidance Clinical Data Presentations for Orthopedic Device Applications (2004), using the following surgical categories: revisions, removals, reoperations, and other, as applicable. This table may be generated manually.

12.12. Other Data

Protocol deviations will be listed and summarized. Surgical information will be summarized. Individual components of the Functional Comorbidity Index (FCI) questionnaire may be summarized.

13. Comparability to Aequalis Dataset

Demographics and baseline characteristics for subjects in the Aequalis dataset and subjects with the investigational system will be compared by t-tests for continuous factors and chi square tests for categorical factors; non-parametric or exact tests may be used when appropriate. These characteristics will include age, gender, diagnosis, and baseline adjusted Constant score. No imputations will be done for these comparisons. Additional baseline measures may be considered if sufficient non-missing values are available. A propensity score analysis may also be performed on the subjects with complete 24 month status/data available if differences in the populations are observed. The propensity score analysis will incorporate data from Aequalis dataset used to derive the performance goal and subjects enrolled in the current investigation. If significant differences are found between groups, the impact of relevant factor(s) on the primary endpoint success rate will be assessed by logistic regression.

13.1. *Aequalis Data*

A subset of data from the Tornier Aequalis Post-Market Outcomes Study was used to develop the performance goal for the primary endpoint hypothesis test of the current investigation. A total of 191 subjects from the post-market study had an indication similar to that of the current investigation and had endpoint status at 24 months determined. Their available data will be used in the comparison to the current investigation.

13.2. *Propensity Score Analysis*

A propensity score adjusted analysis will be performed as a supplemental analysis to the primary analysis *if* any of the standardized mean difference between the two groups for the list of confounders differs by more than 10% (see below more specifics). This section provides additional details for that analysis.

Propensity score based methods may be used to account for significant imbalances in baseline confounding characteristics between the current investigation and Aequalis. The final analysis will be a propensity score adjusted analysis on the primary endpoint by using a logistic model to model probability of success adjusted by the final propensity score. 95% confidence bounds will be calculated for the difference in probability of success between the current investigation and Aequalis (Aequalis – current investigation). If the upper bound for the mean difference excludes the non-inferiority margin of 10%, non-inferiority has been met for the propensity adjusted analysis.

The current investigations study data will be sent from the sponsor to the independent statistician in the format of SAS datasets. The sponsor does not have capability to access or analyze data in this format.

These analyses will be performed by two statisticians, both independent statisticians (i.e. not sponsor employees):

- The propensity score statistician (PS statistician) will receive blinded data from the independent statistician and will construct the propensity scores and have no access to the outcomes of the trial to prevent bias. The PS statistician will remain blinded to the outcomes.
- The independent statistician is unblinded and will prepare a blinded SAS dataset with the specified covariates and study ID with no outcome data and send to the PS statistician. The unblinded statistician will complete all other summaries for the report and incorporate the final propensity scores received from the PS statistician into the primary endpoint comparison. The PS statistician will not complete any analysis that involves the study endpoints.

The propensity score is the likelihood that a patient would be assigned to the current investigation given the profile of his/her baseline characteristics. The propensity score will be derived from a propensity score model developed using a multivariate logistic regression with the study (current investigation vs Aequalis data) as the response variable and the following potential baseline confounders as predictors:

- Gender
- Age

- Baseline adjusted constant score
- Diagnosis
 - Avascular Necrosis
 - Primary Glenohumeral Osteoarthritis
 - Post-Traumatic Arthritis
- Body Mass Index

Simple imputation will be done to any missing baseline data points used in the propensity score model. For continuous covariates with missing information, the mean population value for the study (i.e., Aequalis data, current investigation) will be used. For categorical covariates with missing data, the level with the highest proportion of patients for the study will be assigned.

Building the Propensity Model (performed by blinded propensity score statistician)

The propensity score is a probability (of being in the current investigation) and will range from 0 to 1, such that patients with a higher score are more likely to have been part of the current investigation than those with lower propensity scores.

The fit of the propensity score model is critical. All confounders will be entered into the logistic regression model and kept in the model regardless of statistical significance due to the small number of confounders. One logistic model will use the confounders only as main effects. The second logistic model will use the main effects, the quadratic terms for age and baseline adjusted constant score and all 2-way interaction terms.

Checking for improvement in confounder balance between the original (unadjusted) comparison and the propensity score adjusted comparison

The imbalance in confounders will be assessed in the original study groups. Per Austin (2006) the standardized mean difference will be used and the sum of the absolute value of the standardized differences over all confounders will serve as the overall measure of covariate balance in the original group.

Per Austin (2008), the equivalent to the absolute value of the standardized mean difference when covariate adjustment is used is the absolute value of the weighted conditional standardized difference. This value will be summed over all confounders and compared to the value in the original group for the two proposed propensity score models. The propensity score model with the lowest sum of the absolute values of the weighted conditional standardized difference will be selected. It is expected that this sum will be less than that calculated for the original group and if so represents the improvement in covariate balance due the adjustment by the propensity score in the primary endpoint comparison.

However, if any of the individual absolute values of the weighted conditional standardized differences are greater than 10% (0.10) with the selected propensity model, we will then follow the “rule of thumb” suggestion by Crump et al. 2009, and use only propensity scores in the range [0.1, 0.9] to see if that resolves the issue. If after limiting the subjects to that range, there are no mean differences greater than 10% and the summed value has decreased further, only those subjects with propensity scores in that range will be used in the comparison. If it does not, we will limit the sample to the region of common support and if that resolves the issue, the

patients in the region of common support will be used. If it does not, the full sample size in both groups will be used.

Using the selected propensity score model, the extent of overlap between the study groups will be compared using box plots. Additionally, the propensity scores for all subjects (current investigation and Aequalis, combined) will be stratified into quintiles (each containing as close to 20% of the population as possible) and the proportion of Pyrocarbon and Aequalis subjects in each quintile will be compared to see if there is sufficient overlap in the scores. Sufficient overlap will be defined as at least 10 subjects of each group (current investigation vs. Aequalis) within a stratum. If sufficient overlap is not found, we will limit the primary endpoint comparison using either “rule of thumb” suggestion by Crump et al. 2009, [0.1, 0.9] or the region of common support.

Ultimately, the largest sample of patients providing improved balance in confounders with acceptable overlap will be used in the analysis. As a sensitivity analysis, the propensity score adjusted primary endpoint comparison will be made in the remaining two patient groups as appropriate, i.e. sufficient overlap exists and improvement in confounder balance was achieved.

14. Version History

Version	Date	Changes
1.0	20JUL2015	Initial version.
2.0	17AUG2015	Updated to include 150 total enrollments Clarified difference between enrolled and implant attempts
3.0	18AUG2015	Clarified unsuccessful implant attempts are primary endpoint failures
4.0	11SEP2015	Updated to match updated protocol including updated sample size calculation, data collection, follow-up windows, adding revision rate and radiographic summaries comparing to Aequalis
5.0	18AUG2017	Updated and expanded propensity score adjusted analysis.
6.0	07JAN2019	Updated to add BMI to propensity score model. Added a revision rate for secondary endpoint.

15. References

1. Guidance for Industry and FDA Staff: Clinical Data Presentations for Orthopedic Device Applications (2004);
<http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm072263.htm>
2. Michener, PT. et al. Understanding the relevance of measured change through studies. Spine 2000; 25:3192-9
3. Constant C R, et al. A review of the Constant Score: Modifications and guidelines for its use. J Shoulder Elbow Surg 2008; 17(2): 355-261
4. Johansson, KM. et al. Intraobserver and interobserver reliability for the strength test in the Constant-Murley shoulder assessment. J Shoulder Elbow Surg. 2005;14:273–278
5. C. Constant, "Age related recovery of shoulder function after injury [MCh Thesis]," University College, Cork, Ireland, 1986.

6. <http://www.orthop.washington.edu/?q=patient-care/articles/shoulder/simple-shoulder-test.html>
7. Romeo, A et al. Shoulder Scoring Scales for the Evaluation of Rotator Cuff Repair. Clinical Orthopaedics and Related Research. 2004; 427:107-114
8. Austin PC and MM Mamdani, A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. Statistics in Medicine 2006; 25: 2084-2106.
9. Crump RK, VJ Hotz, GW Imbens and OA Mitnik. Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96. 187-199.
10. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf. 2008; 17: 1202-1217.