http://dx.doi.org/10.2106/JBJS.20.02002

Statistical Analysis Plan (SAP)
Incidence and Risk of Nail Breakage in Patients Implanted with the DePuy Synthes TFN-ADVANCED™ Proximal Femoral Nailing System (TFNA)
Protocol Version: 2

http://dx.doi.org/10.2106/JBJS.20.02002

Table of Contents

http://dx.doi.org/10.2106/JBJS.20.02002

http://dx.doi.org/10.2106/JBJS.20.02002

1. Study Rationale and Design

TFN-Advanced™ Proximal Femoral Nailing System is intended for treatment of fractures in adults and adolescents in which the growth plates have fused. Based on an analysis of the Sponsor's complaint database, an increased number of TFNA nail breakage complaints from Western Australia/Australia were identified and evaluated. Following a July 2019 Product Safety Committee meeting, the team agreed to the execution of a retrospective epidemiology study to describe and compare nail breakage of TFNA to other nails. The current study is a retrospective cohort study using the premier healthcare database in which the primary objective is to estimate the difference in risk of nail breakage for TFNA vs. comparable nail systems. The study has only one confirmatory analysis (see 4.1-4.8), all other analyses are exploratory.

1.1 Primary Objective

The primary objective is to estimate the incidence rate and the relative risk of nail breakage among patients implanted with the DePuy Synthes TFN-ADVANCED™ Proximal Femoral Nailing System (TFNA) compared to patients implanted with selected nails designed to provide cephalomedullary support to a proximal femoral fracture, the Stryker® Gamma3® Nail System or Zimmer® Natural Nail® System. ***Determinations about the safety of TFNA will be based on the risk difference of nail breakage between these two groups using data balanced on measured covariates.***

1.2 Secondary Objectives

The secondary objectives include the following:

1. To describe the demographic and clinical characteristics of patients implanted with DePuy Synthes TFNA, DePuy Synthes TFN, Stryker Gamma3, and Zimmer Natural Nail Systems
2. To estimate the mean and median time to nail breakage among patients implanted with TFNA, Stryker Gamma3, and Zimmer Natural Nail who have nail breakage
3. To estimate and compare the incidence rate of nail breakage by diameter of cephalomedullary nail (~15mm and ~17mm)
4. To estimate nail breakage incidence rates by calendar year stratified by diameter of cephalomedullary nail (~15mm or ~17mm)
5. To estimate nail breakage incidence rates stratified by length of cephalomedullary nail (short or long)
6. To estimate nail breakage incidence rates stratified by patient characteristics including age, sex, race, type of fracture, obesity diagnosis, presence of pathologic fracture, and presence of polytrauma

http://dx.doi.org/10.2106/JBJS.20.02002

2. Analysis Data Sets

The study population consists of patients with femur fracture undergoing surgical repair with a cephalomedullary nail system in an inpatient setting. The primary objective and secondary objective 2 are evaluated on:

- Patients ≥21 years old who have an ICD-9/10 procedure code for femur fracture repair with internal fixation device
- Patients surgically treated with DePuy Synthes TFNA, Stryker Gamma 3 or Zimmer Natural Nail cephalomedullary nail in an inpatient setting between February 1, 2014 and September 30, 2019
- Patients who do not have missing data on age or sex (an indicator of poor data quality)
- Patients who did not experience a nail breakage on the index procedure date
- Patients who did not experience bilateral femur fractures repaired with bilateral cephalomedullary nails, defined based on the presence of bilateral procedures (for ICD-10, both right and left procedure codes present; for ICD-9, two procedure codes present) and billing charges for 2 or more nails occurring during the index episode of care

Secondary objectives 1,3,4,5, and 6 will be evaluated on:

- Patients ≥21 years old who have an ICD-9/10 procedure code for femur fracture repair with internal fixation device
- Patients surgically treated with DePuy Synthes TFN, DePuy Synthes TFNA, Stryker Gamma 3 or Zimmer Natural Nail cephalomedullary nail in an inpatient setting between January 1, 2010 and September 30, 2019
- Patients who do not have missing data on age or sex (an indicator of poor data quality)
- Patients who did not experience a nail breakage on the index procedure date
- Patients who did not experience bilateral femur fractures repaired with bilateral cephalomedullary nails, defined based on the presence of bilateral procedures (for ICD-10, both right and left procedure codes present; for ICD-9, two procedure codes present) and billing charges for 2 or more nails occurring during the index episode of care

3. Sample Size Justification

http://dx.doi.org/10.2106/JBJS.20.02002

Study design was informed by sample size considerations, which are based on statistical power. Power was calculated by Monte Carlo simulation using 5,000 simulated datasets. Of primary interest is power to identify a safety success (i.e, presence of a superiority, non-inferiority or equivalence study outcome, see 4.7). This equates to power for a conventional test of non-inferiority with a margin of non-inferiority of 0.5% for the difference in the percent cumulative failure (PCF) between the two treatment arms at 18 months. The estimate of the PCF at 18 months uses a weighted estimate of the survival probability to induce group balance (see 4.6 and 4.8), $\widehat{PCF} = [1 - \hat{S}(18)] * 100$, with the difference estimated as $\hat{\Delta} = \widehat{PCF}_{TFNA} - \widehat{PCF}_{Non-TFNA}$. The standard error of the difference is calculated as $\widehat{SE}_{\hat{\Delta}} = 100 * \sqrt{\widehat{\text{var}}\,(\widehat{PCF}_{TFNA}) + \widehat{\text{var}}\,(\widehat{PCF}_{Non-TFNA})}$, where each variance term allows for dependency among observations within the same cluster (hospital) [1,2]. The two-sided 95% confidence interval used for the hypothesis test is $\hat{\Delta} \pm Z_{1-\alpha/2} * \widehat{SE}_{\hat{\Delta}}$. If the upper bound of the 95% confidence interval for $\hat{\Delta}$ is smaller than 0.5% then it is counted as a safety success (4.7). The proportion of safety success across the 5,000 simulations is the empirical estimate of power. In these simulations, the population difference between the two treatment arms is set to Δ=0.0%.

The sample size for the analysis was based on the number of observations in the covariate balanced dataset. For each participant a time-to-event outcome was generated using the method of Bender et al. [3]. The event time was generated by selecting a random number from a standard uniform distribution, $u_{ck} \sim U(0,1)$, with C sites $(c = 1, \ldots, C)$, n_c observations per site $(k = 1, \ldots, n_c)$ and $N = \sum_{c=1}^{C} n_c$ all based on the data. This quantity is then used to compute: $\left(\frac{-\log(u_{ck})}{\lambda \exp(\gamma_c)}\right)^{1/\eta}$ with $\lambda = .15$, $\eta = 1$, $\gamma_c \sim N(0, \sigma_\gamma^2)$. Between-site variance, σ_γ^2, was set to 0 or 0.53. The value of 0 corresponds to a median hazard ratio of 1 and 0.53 corresponds to a median hazard ratio of 2. The median hazard ratio is the median relative change in the hazard of the occurrence of the outcome when comparing identical subjects from two randomly selected different clusters that are ordered by risk [4] (see also [5]). Whether a subject experienced an event was randomly sampled $E_{ck} \sim Bernoulli\,(p_{ck})$ with the event probability, p_{ck}, set to 0.0448 or 0.0224. With p_{ck}= 0.0448 the PCF is 1.0% at 18 months, whereas with p_{ck}= 0.0224 the PCF is 0.5% at 18 months.

As can be seen in Table 1, power is generally adequate (>0.80) with lower event proportions but not with higher event proportions. Use of cluster robust standard errors is particularly useful for accurately estimating the standard errors in the presence of between-cluster variability. When examining the ratio of the mean standard error to the empirical standard error $\left(\overline{\widehat{SE}}\left(\hat{S}(18)\right)/\widehat{SD}\left(\hat{S}(18)\right)\right)$ we obtain values near 1 for the conditions evaluated in Table 1 using cluster robust standard errors: 0.96 (PCF=1.0%, $\sigma_\gamma^2 = 0$), 0.96 (PCF=1.0%, $\sigma_\gamma^2 = 0.53$), 0.97 (PCF=0.5%, $\sigma_\gamma^2 = 0$), 0.95 (PCF=0.5%, $\sigma_\gamma^2 = 0.53$). By comparison, standard errors that do not account for clustering produced the following ratios: 0.98, 0.80, 0.99 and 0.87.

http://dx.doi.org/10.2106/JBJS.20.02002

Table 1. Power for a Test of Non-Inferiority Under Varying PCF and Cluster Variability

PCF at 18 months	No Cluster Variability ($\sigma_\gamma^2 = 0.00$)	Cluster Variability ($\sigma_\gamma^2 = 0.53$)
1.0%	0.67	0.40
0.5%	0.97	0.83

Another consideration is power for a test for inferiority (see 4.7). For this test we use the same simulation parameters described above with the following exception. In these simulations, the population difference between the two treatment arms is set to Δ=1.0%. This was done by setting p_{ck}= 0.0672 in the TFNA arm, corresponding to a PCF of 1.5% at 18 months, and setting p_{ck}= 0.0224 in the Non-TFNA arm, corresponding to a PCF of 0.5% at 18 months. If the lower bound of the 95% confidence interval for $\hat{\Delta}$ is larger than 0.5%, which is equivalent to the methods for establishing inferiority described in 4.7, then a safety signal is present. The proportion of safety signals across the 5,000 simulations is the empirical estimate of power. Power is 0.79 in the presence of between-cluster variability ($\sigma_\gamma^2 = 0.53$). Moreover, use of cluster robust standard errors is particularly useful for accurately estimating the standard errors in the presence of between-cluster variability. When examining the ratio of the mean standard error to the empirical standard error ($\overline{\widehat{SE}}\left(\hat{S}(18)\right)/\widehat{SD}\left(\hat{S}(18)\right)$) we obtain 0.97, whereas standard errors that do not account for clustering produced a value of 0.77.

4. Analytic Methods to Meet the Primary Objective

4.1. Treatment

The treatment of interest consists of patients receiving DePuy Synthes TFN-ADVANCED™ Proximal Femoral Nailing System (TFNA) for a proximal femoral fracture. The comparison group is patients with use of either a Stryker® Gamma3® Nail System or Zimmer® Natural Nail® System for a proximal femoral fracture.

4.2 Covariates

Covariates used in calculation of the propensity score include: age at index procedure (continuous), sex (nominal: male/female), race (nominal: White, Black, Other, Unknown), calendar year of index procedure (nominal: 2014/2015,2016,2017,2018,2019), Nail length (nominal: short(≤ 235 mm), long(>235 mm), unknown), fracture type (nominal, not mutually exlusive categories: subtrochanteric, pertrochanteric (all trochanteric not categorized as subtrochanteric), intracapsular (inclusive of femoral head and neck), femoral shaft, other fractures), pathological fracture (nominal: yes,no), bone neoplasm (nominal: yes,no), injury severity score (nominal: <15, ≥15), polytrauma (nominal: yes, no), summary score of Elixhauser comorbidities recorded during or prior to the index procedure (continuous), individual Elixhauser comorbidities (obesity, solid tumor without metastasis, metastatic cancer,

lymphoma; all nominal: yes, no), additional ICD-9/10 comorbidities (dementia, syncope, osteoporosis/osteopenia; all nominal: yes, no), surgical specialty (nominal: orthopedic surgeon, general surgeon, other), hospital bed size (nominal: 0-199, 200-299-399, 400+), hospital teaching status (nominal: teaching, non-teaching), hospital setting (nominal: rural, urban), hospital region (nominal: midwest, northeast, south, west).

4.3 Endpoint Creation

The outcome of interest for the primary objective is time to nail breakage after the hospitalization in which the index procedure was performed. For patients experiencing a nail breakage, defined as a subsequent hospitalization in which *both* a diagnosis of breakdown of internal device and procedure for femur fracture repair/device removal from femur are present, the event time is calculated as time from the index procedure hospitalization to the re-hospitalization used to treat the nail breakage. Other censoring events include end of hospital participation in the premier healthcare database or the last month in which a breakage could be recorded in the database (September, 2019). Among these censoring events, calculation of time will be based on whichever censoring event occurs first. Since exact dates are not available for any hospitalization, time is calculated as the number of months elapsed from the index procedure hospitalization to the outcome event or a censoring event. Therefore, time will take on the following values: 0.5 (outcome event/censoring event occurs in the same calendar month as the index procedure hospitalization; i.e. actual time to event must be <1 month), 1 (month/year of outcome event/censoring event – month/year of index procedure hospitalization = 1; i.e. outcome event/censoring event occurs in the calendar month following index procedure hospitalization),…, 18 (month/year of outcome event/censoring event – month/year of index procedure hospitalization = 18 months). Intraoperative and post-operative breakage during the initial hospital stay are not included as events or censored cases because of the inability to use billing records to distinguish between primary and revision procedures.

4.4 Missing Data

Missing data are only present on the covariates of race and nail length. We create a separate level to indicate missing for these nominal variables [6]. The variables with levels for missing data will be included in the propensity score model.

4.5 Observational Study Conduct

Covariate balancing methods are used to address the comparisons of TFNA to non-TFNA nails as part of the primary objective. In our approach, covariate balancing is kept separate from the analysis [7]. A physical separation between the design and analysis is put in place [8,9], such that the person responsible for covariate balancing does not have access to the outcome data. Specifically, one individual who is a member of a contract organization (MuSigma) created the analytical dataset. Information on the outcome variable was removed from the analytical dataset by study biostatistician 1 while retaining a linkage id, and this outcome-removed dataset was provided to study biostatistician 2. The outcome-removed dataset was used to balance the data using weights (4.6) by study biostatistician 2. The outcome-removed dataset with additional

columns for weights will be provided to study biostatistician 1, who will remerge it to the outcome variables by using the linkage id and perform an outcome analysis (4.8).

4.6 Covariate Balancing

The propensity score is used for covariate balancing to estimate the average treatment effect on the treated. The propensity score is calculated using a multivariable logistic regression model, with K-1 dummy variables for nominal covariates with K levels, and a single variable for each continuous covariate. We initially considered several methods of balancing the data: 1) nearest neighbor 1:1 matching with a caliper (i.e., $0.20 * SD_{logit(PS)}$) and sampling controls with replacement, 2) stratification on the propensity score using 5 or 10 strata, 3) optimal full matching [10,11], where treated cases receive a weight of 1 and controls receive a weight proportional to the number of treated cases divided by the number of controls in the matched set, $\frac{|A_m|}{|B_m|}$, where $A = treated\ units, B = control\ units, and\ m = 1, \ldots, M\ matched\ sets$, and 4) average treatment effect on the treated weights, where treated cases receive a weight of 1 and untreated receive a weight based on the odds, $\frac{e_i}{1-e_i}$, where $e_i = \text{propensity score}$. With the use of weights we also considered weight trimming [12] at the 99.9, 99.5, 99th, 98th, or 97th. When used with logistic regression to estimate the propensity score, weight trimming has been shown to reduce bias and increase precision of the estimated treatment effect [12]. However, since balancing is done independent of the outcome data, trimming is only undertaken if it does not worsen balance relative to the untrimmed weights.

Research suggests that most of the above covariate balancing methods are effective (i.e., unbiased) in estimating the proposed risk difference [13]. Among those methods shown to be effective include caliper matching (without replacement) and ATT weights. Optimal matching was not evaluated in [13], however the method has been shown effective in estimating the marginal hazard ratio [11]. Notably, stratification on the propensity score performed sub-optimally in estimating the risk difference for survival data [13], which is consistent with previous research for estimating differences in proportions and means [14,15].

For each covariate, the mean and standard deviation (continuous variables) or proportion and percentage (nominal covariates) is calculated for each treatment group. Balance is evaluated using absolute standardized differences [16]. For continuous covariates we use $\hat{d} = \frac{|\bar{x}_t - \bar{x}_c|}{\sqrt{\frac{\hat{v}(\bar{x}_t) + \hat{v}(\bar{x}_c)}{2}}}$ and for binary covariates $\hat{d} = \frac{|\hat{p}_t - \hat{p}_c|}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_c)}{2}}}$. Nominal variables with more than two categories are summarized using a generalization of the absolute standardized difference for binary covariates [17].

Among the covariate balancing methods described above the one that minimized imbalance was average treatment effect on the treated weights. Our implementation of the weights involves trimming at the 98th percentile of the distribution of weights in the comparison group, as this did not result in an increase in imbalance. We evaluated

http://dx.doi.org/10.2106/JBJS.20.02002

balance for each method using the number of variables with absolute standardized differences (NASD) less than 0.10, as well as the average absolute standardized differences (AASD) across the 27 covariates. Matching (NASD=5, AASD=0.081), optimal full matching (NASD=6, AASD=0.089), stratification with 5 strata (NASD=7, AASD=0.085) and stratification with 10 strata (NASD=10, AASD=0.104) were all inferior to average treatment effect on the treated weights without trimming (NASD=2, AASD=0.032) or with trimming at the 98th percentile (NASD=2, AASD=0.032).Table 2 displays the data before balancing and Table 3 displays the data after balancing using average treatment effect on the treated weights with trimming at the 98th percentile.

Table 2. Covariate Balance in Unbalanced Data

		Non-TFNA	TFNA	ASD
n		8260	14370	
Gender (Male)		2416 (29.2)	4478 (31.2)	0.042
Race				**0.218**
	Black	339 (4.1)	544 (3.8)	
	Other	944 (11.4)	799 (5.6)	
	Unknown	80 (1.0)	215 (1.5)	
	White	6897 (83.5)	12812 (89.2)	
Year				**1.048**
	2016	1606 (19.4)	2642 (18.4)	
	2017	1442 (17.5)	3825 (26.6)	
	2018	956 (11.6)	3770 (26.2)	
	2019	677 (8.2)	3289 (22.9)	
	2014/2015	3579 (43.3)	844 (5.9)	
Nail Length				**0.437**
	Long	3333 (40.4)	7777 (54.1)	
	Short	4702 (56.9)	5360 (37.3)	
	Unspecified	225 (2.7)	1233 (8.6)	
Subtrochanteric Fracture (Yes)		753 (9.1)	1563 (10.9)	0.059
Pertrochanteric Fracture (Yes)		6637 (80.4)	11355 (79.0)	0.033
Intracapsular Fracture (Yes)		362 (4.4)	514 (3.6)	0.041
Femoral Shaft Fracture (Yes)		146 (1.8)	363 (2.5)	0.052
Other Fracture (Yes)		73 (0.9)	145 (1.0)	0.013
Pathological Fracture (Yes)		601 (7.3)	1031 (7.2)	0.004
Bone Neoplasm		202 (2.4)	449 (3.1)	0.041
Injury Severity Score (>15)		8166 (98.9)	14074 (97.9)	0.074
Polytrauma (Yes)		3197 (38.7)	2030 (14.1)	**0.580**
Lymphoma (Yes)		107 (1.3)	271 (1.9)	0.047
Metastatic Cancer (Yes)		323 (3.9)	723 (5.0)	0.054
Solid Tumor Without Metastasis (Yes)		717 (8.7)	1558 (10.8)	0.073
Obesity (Yes)		753 (9.1)	1820 (12.7)	0.114
Dementia (Yes)		1187 (14.4)	2244 (15.6)	0.035
Osteoporosis/Osteopenia (Yes)		2873 (34.8)	5259 (36.6)	0.038
Syncope (Yes)		629 (7.6)	1404 (9.8)	0.077

http://dx.doi.org/10.2106/JBJS.20.02002

Surgical Specialty				**0.290**
	General	37 (0.4)	53 (0.4)	
	Orthopedic	7151 (86.6)	13623 (94.8)	
	Other	1072 (13.0)	694 (4.8)	
Number of Beds				**0.244**
	000-199	1947 (23.6)	2036 (14.2)	
	200-399	3001 (36.3)	6084 (42.3)	
	400+	3312 (40.1)	6250 (43.5)	
Teaching Hospital (Yes)		2629 (31.8)	5350 (37.2)	**0.114**
Hospital Setting (Urban)		6967 (84.3)	12043 (83.8)	0.015
Hospital Region				**0.302**
	Midwest	1584 (19.2)	2843 (19.8)	
	Northeast	1258 (15.2)	907 (6.3)	
	South	3860 (46.7)	7979 (55.5)	
	West	1558 (18.9)	2641 (18.4)	
Age		78.43 (11.53)	77.73 (12.11)	0.059
Elixhauser Index Score (sum of comorbidities)		4.68 (2.95)	5.18 (3.28)	**0.159**

Note: ASD=absolute standardized difference. Bold indicates a value above 0.100.

Table 3. Covariate Balance in Balanced Data

		Non-TFNA	TFNA	STD
n		13878.3	14370	
Gender (Male)		4267.8 (30.8)	4478.0 (31.2)	0.009
Race				0.023
	Black	539.3 (3.9)	544.0 (3.8)	
	Other	831.4 (6.0)	799.0 (5.6)	
	Unknown	226.6 (1.6)	215.0 (1.5)	
	White	12281.0 (88.5)	12812.0 (89.2)	
Year				0.061
	2016	2469.1 (17.8)	2642.0 (18.4)	
	2017	3393.0 (24.4)	3825.0 (26.6)	
	2018	3900.3 (28.1)	3770.0 (26.2)	
	2019	3315.1 (23.9)	3289.0 (22.9)	
	2014/2015	800.9 (5.8)	844.0 (5.9)	
Nail Length				0.097
	Long	7342.4 (52.9)	7777.0 (54.1)	
	Short	5643.3 (40.7)	5360.0 (37.3)	
	Unspecified	892.5 (6.4)	1233.0 (8.6)	
Subtrochanteric Fracture (Yes)		1491.1 (10.7)	1563.0 (10.9)	0.004
Pertrochanteric Fracture (Yes)		11006.0 (79.3)	11355.0 (79.0)	0.007
Intracapsular Fracture (Yes)		560.6 (4.0)	514.0 (3.6)	0.024

http://dx.doi.org/10.2106/JBJS.20.02002

Femoral Shaft Fracture (Yes)		317.3 (2.3)	363.0 (2.5)	0.016
Other Fracture (Yes)		142.0 (1.0)	145.0 (1.0)	0.001
Pathological Fracture (Yes)		936.3 (6.7)	1031.0 (7.2)	0.017
Bone Neoplasm		382.6 (2.8)	449.0 (3.1)	0.022
Injury Severity Score (>15)		13627.6 (98.2)	14074.0 (97.9)	0.018
Polytrauma (Yes)		1634.4 (11.8)	2030.0 (14.1)	0.070
Lymphoma (Yes)		265.9 (1.9)	271.0 (1.9)	0.002
Metastatic Cancer (Yes)		610.1 (4.4)	723.0 (5.0)	0.030
Solid Tumor Without Metastasis (Yes)		1450.6 (10.5)	1558.0 (10.8)	0.013
Obesity (Yes)		1578.3 (11.4)	1820.0 (12.7)	0.040
Dementia (Yes)		2128.4 (15.3)	2244.0 (15.6)	0.008
Osteoporosis/Osteopenia (Yes)		4740.3 (34.2)	5259.0 (36.6)	0.051
Syncope (Yes)		1330.6 (9.6)	1404.0 (9.8)	0.006
Surgical Specialty				0.013
	General	49.6 (0.4)	53.0 (0.4)	
	Orthopedic	13119.0 (94.5)	13623.0 (94.8)	
	Other	709.7 (5.1)	694.0 (4.8)	
Number of Beds				**0.122**
	000-199	2594.0 (18.7)	2036.0 (14.2)	
	200-399	5564.5 (40.1)	6084.0 (42.3)	
	400+	5719.7 (41.2)	6250.0 (43.5)	
Teaching Hospital (Yes)		5309.6 (38.3)	5350.0 (37.2)	0.021
Hospital Setting (Urban)		11696.6 (84.3)	12043.0 (83.8)	0.013
Hospital Region				**0.115**
	Midwest	2273.6 (16.4)	2843.0 (19.8)	
	Northeast	1114.1 (8.0)	907.0 (6.3)	
	South	7631.4 (55.0)	7979.0 (55.5)	
	West	2859.2 (20.6)	2641.0 (18.4)	
Age		77.92 (11.61)	77.73 (12.11)	0.015
Elixhauser Index Score (sum of comorbidities)		5.02 (3.08)	5.18 (3.28)	0.049

Note: ASD=absolute standardized difference. Bold indicates a value above 0.100.

4.7 Hypothesis Testing for Comparison of TFNA to Non-TFNA Nails in Covariate Balanced Data

As described in 1.1, the primary analysis used to make safety determinations in this study is comparison of TFNA to Non-TFNA nails (i.e., Gamma3 and Natural Nail). For this comparison the estimate of interest is the risk difference of nail breakage between these two groups using data balanced on measured covariates. From a safety perspective, it is of primary interest to demonstrate that TFNA is as safe as Non-TFNA nails with respect to nail breakage. Given this objective, a test of non-inferiority is appropriate. It may be also of interest to test for superiority (TFNA has lower risk for nail breakage than non-TFNA nails) and inferiority (TFNA has higher risk for nail breakage than non-TFNA nails). First, denote the treatment effect of interest as the difference in

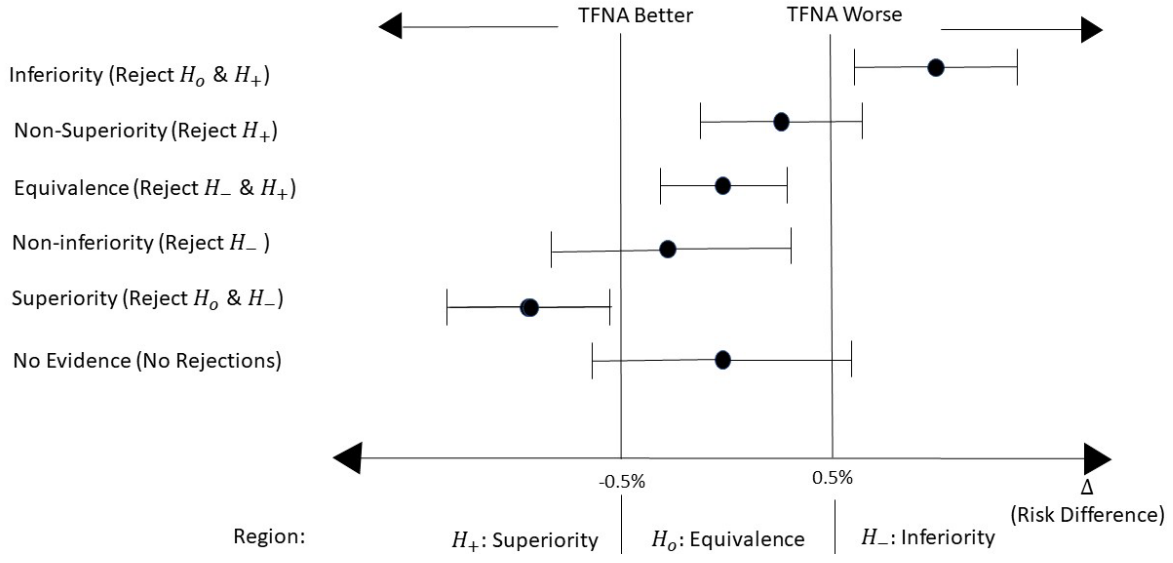
http://dx.doi.org/10.2106/JBJS.20.02002

the percent cumulative failure (PCF) between the two treatment arms at 18 months $\Delta = PCF_{TFNA} - PCF_{Non-TFNA}$, where $PCF = [1 - S(18)] * 100$. To simultaneously test these three hypotheses, we partition the parameter space into three disjoint regions based on a margin of equivalence of 0.5%:

$H_+: \Delta < -0.5\%$ (superiority)
$H_o: -0.5\% \leq \Delta \leq 0.5\%$ (equivalence)
$H_-: \Delta > 0.5\%$ (inferiority)

Three tests are performed, each test used to determine whether the parameter Δ is in the domain of the region. The null hypothesis for the test of H_+ is $\Delta \in H_+$ and the alternative hypothesis is $\Delta \in H_+^c$ (the complement of H_+), the null hypothesis for a test of H_o is $\Delta \in H_o$ and the alternative hypothesis is $\Delta \in H_0^c$, and the null hypothesis for a test of H_- is $\Delta \in H_-$ and the alternative hypothesis is $\Delta \in H_-^c$. Using a two-sided 95% confidence interval for $\hat{\Delta}$ (4.8), we reject a null hypothesis if the domain of that null hypothesis does not overlap with the confidence interval. The approach described here is based on the partitioning principle and therefore allows for simultaneous testing of equivalence, superiority and inferiority with strong control of the familywise error rate at the nominal α level [18,19]. Note that the approach to hypothesis testing taken here is not the same as what is proposed by Goemann et al. [19], in order that hypothesis testing be consistent with how 'classical' confidence intervals are calculated. Figure 1 displays possible outcomes from these tests and their interpretations. *A safety success in this study would be demonstrated if either superiority, non-inferiority or equivalence is present. However, distinguishing among these three outcomes can also be informative. In contrast, a safety signal would be demonstrated if inferiority is present. If either non-superiority or no evidence is present, this would indicate an inconclusive study result.*

Figure 1. Inferences Under Possible Study Outcomes



http://dx.doi.org/10.2106/JBJS.20.02002

4.8 Outcome Analysis for Comparison of TFNA to Non-TFNA Nails in Covariate Balanced Data

The approach to outcome modeling in this study is to estimate difference in the percent cumulative failure (PCF) at 18 months between TFNA and Non-TFNA groups. Given the method used to balance the data, survival estimates are weighted according to average treatment effect on the treated weights (4.6). Survival is calculated using the Nelson-Aalen estimator. The method applies weights to calculation of the number at risk and the number of events at each distinct event time, as with the weighted estimator of survival using Kaplan-Meier estimator described elsewhere [20]. With respect to variance estimation within each treatment arm, we allow for dependency among observations on the outcome within the same cluster (hospital) [1,2]. Simulations indicate that in the presence of dependency, the method more accurately estimates the standard error of survival probabilities when compared with conventional methods that assume independence, even with small cluster sizes and rare events [21] (see also simulation results reported in section 3). A point estimate of the difference is based on: $\hat{\Delta} = \widehat{PCF}_{TFNA} - \widehat{PCF}_{Non-TFNA}$, where $\widehat{PCF} = [1 - \hat{S}(18)] * 100$. The standard error of the difference is calculated as $\widehat{SE}_{\hat{\Delta}} = 100 * \sqrt{\widehat{\text{var}}\,(\widehat{PCF}_{TFNA}) + \widehat{\text{var}}\,(\widehat{PCF}_{Non-TFNA}))}$ and the 95% confidence interval is $\hat{\Delta} \pm Z_{1-\alpha/2} * \widehat{SE}_{\hat{\Delta}}$

4.9 Supplementary Analyses

4.9.1 Sensitivity Analysis Using A Negative Control Outcome

A sensitivity analysis will be performed for the confirmatory analysis described in 4.1-4.8. Specifically, we will test the possibility of residual confounding using a negative control outcome [26]. The outcome in this case will be time to cataract diagnosis. No hypothesis testing will be performed for this analysis.

4.9.2 Sensitivity Analysis Restricting to ICD-10 Data

A sensitivity analysis will be performed for the confirmatory analysis described in 4.1-4.8. Specifically, analyses will be re-executed using data from October 2015 through September 2019 only to evaluate whether the results are sensitive to changes in coding of exposures and outcomes that occurred with the transition to ICD-10.

4.9.3 Hazard Ratio from a Cox Proportional Hazards Model

One analysis used to support the results from the confirmatory analysis is a comparison of TFNA and Non-TFNA in the covariate balanced data that estimates the hazard ratio using a weighted Cox model. The model is relatively unbiased with a higher prevalence of those receiving treatment [22]. We allow for dependency among observations within the same cluster (hospital) by cluster robust standard errors [23] using a finite cluster bias correction of C/(C-1), which accurately estimate the standard errors given the clustering and balancing design [24].The results of this analysis will not be used to judge the safety of TFNA, which is based on the analysis described in 4.8, rather the intention is to characterize the treatment effect using a hazard ratio.

http://dx.doi.org/10.2106/JBJS.20.02002

4.9.4 Subgroup Analyses

Various subgroup analyses will be performed. One analysis will include patients that have pertrochanteric fractures (allowing for concurrent fractures of other types). Another analysis will include patients that have either pertrochanteric or subtrochanteric fractures (allowing for concurrent fractures of other types). For these two subgroup analyses, covariate balancing is performed separately for each of these two subgroups of patients and an outcome analysis is executed using the same methods to meet the primary study objective using all fracture types (4.1-4.8).

4.9.5 Risk factors if a Safety Signal is Identified

If a safety signal is present in the primary analysis with all fracture types (i.e., TFNA nails are inferior to Non-TFNA nails with respect to nail breakage), regression analyses will be conducted to determine risk factors for nail breakage. This will be based on a model with main effects for treatment, fracture type, nail length and covariates, as well as terms to represent the interaction between treatment and fracture type as well as treatment and nail length.

4.9.6 Reporting of Percent Cumulative Failure and Incidence Rates

Percent cumulative failures (PCF) and incidence rates of nail breakage by brand over time will be reported. PCF will be reported in the unbalanced (Table Shell 1) and balanced (Table Shell 2) data. Incidence rates will also be reported in the unbalanced (Table Shells 3 & 4) and balanced (Table Shells 5 &6) data.

Table Shell 1. Percent Cumulative Failure and Interval Estimates (95%) at Fixed Times by Nail Brand in the Unbalanced Data

	Time		
Nail Brand	6 months	12 months	18 months
TFNA	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)
Gamma3/ Natural Nail	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)

Note: Percent cumulative failure is based on $\widehat{PCF} = [1 - \hat{S}(t_0)] * 100$. Survival at a specified time, $\hat{S}(t_0)$, is calculated using the Nelson-Aalen estimator. With respect to variance estimation within each treatment arm, we allow for dependency within the same cluster (hospital) [1,2]. Confidence intervals are computed using a log transformation of survival [25]

Table Shell 2. Percent Cumulative Failure and Interval Estimates (95%) at Fixed Times by Nail Brand in the Balanced Data

	Time		
Nail Brand	6 months	12 months	18 months
TFNA	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)
Gamma3/ Natural Nail	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)	XX.XX%(XX.XX,XX.XX)

Note: Percent cumulative failure is based on $\widehat{PCF} = [1 - \hat{S}(t_0)] * 100$. Survival at a specified time, $\hat{S}(t_0)$, is calculated using a weighted Nelson-Aalen estimator. With

http://dx.doi.org/10.2106/JBJS.20.02002

respect to variance estimation within each treatment arm, we allow for dependency within the same cluster (hospital) [1,2]. Confidence intervals are computed using a log transformation of survival [25]

Table Shell 3. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Time and Nail Brand in the Unbalanced Data

	Time		
Nail Brand	0-6 months	0-12 months	0-18 months
TFNA	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Gamma3	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Natural Nail	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

Table Shell 4. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Time and Nail Brand in the Unbalanced Data for Intermediate Intervals

	Time	
Nail Brand	7-12 months	13-18 months
TFNA	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Gamma3	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Natural Nail	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

Table Shell 5. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Time and Nail Brand in the Balanced Data

	Time		
Nail Brand	0-6 months	0-12 months	0-18 months
TFNA	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Gamma3	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Natural Nail	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

Table Shell 6. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Time and Nail Brand in the Balanced Data for Intermediate Intervals

http://dx.doi.org/10.2106/JBJS.20.02002

	Time	
Nail Brand	7-12 months	13-18 months
TFNA	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Gamma3	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
Natural Nail	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

5. Analytic Methods to Meet Secondary Objectives

5.1 Seondary Objective 1

The first secondary objective is to describe the demographic and clinical characteristics of patients implanted with DePuy Synthes TFNA, DePuy Synthes TFN, Stryker Gamma3 and Zimmer Natural Nail Systems. Comparisons among the devices will include all covariates listed in section 4.2 as well as hospital length of stay for the index procedure, malunion and nonunion. Categorical variables will be summarized by frequencies and percentages, and continuous variables will be summarized by means and standard deviations. Differences between study cohorts will be assessed using standardized differences with the reference device being TFNA. See Table Shell 7.

Table Shell 7. Demographic and Clinical Characteristics by Nail Group

	TFNA	TFN	ASD	Gamma3	ASD	Natural Nail	ASD
n	XXX	XXX		XXX		XXX	
Gender (Male)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Race			X.XXX		X.XXX		X.XXX
Black	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Other	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Unknown	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
White	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Year			X.XXX		X.XXX		X.XXX
2016	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
2017	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
2018	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	

http://dx.doi.org/10.2106/JBJS.20.02002

2019	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
2014/2015	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Nail Length			X.XXX		X.XXX		X.XXX
Long	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Short	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Unspecified	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Subtrochanteric Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Pertrochanteric Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Intracapsular Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Femoral Shaft Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Other Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Pathological Fracture (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Bone Neoplasm	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Injury Severity Score (>15)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Polytrauma (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Lymphoma (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Metastatic Cancer (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Solid Tumor Without Metastasis (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Obesity (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Dementia (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Osteoporosis/ Osteopenia (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Syncope (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX

Surgical Specialty			X.XXX		X.XXX		X.XXX
General	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Orthopedic	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Other	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Number of Beds			X.XXX		X.XXX		X.XXX
000-199	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
200-399	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
400+	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Teaching Hospital (Yes)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Hospital Setting (Urban)	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Hospital Region			X.XXX		X.XXX		X.XXX
Midwest	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Northeast	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
South	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
West	XXX (XX.X)	XXX (XX.X)		XXX (XX.X)		XXX (XX.X)	
Age	XX.X (XX.X)	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX
Elixhauser Index Score (sum of comorbidities)	XX.X (XX.X)	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX
Length of Stay	XX.X (XX.X)	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX	XX.X (XX.X)	X.XXX
Malunion	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX
Nonunion	XXX (XX.X)	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX	XXX (XX.X)	X.XXX

5.2 Secondary Objective 2

The second secondary objective is to estimate the mean and median time to nail breakage among patients implanted with TFNA, Stryker Gamma3, and Zimmer Natural

http://dx.doi.org/10.2106/JBJS.20.02002

Nail who have nail breakage. The mean and median time to nail breakage will be reported for each of the devices and across devices. See Table Shell 8.

Table Shell 8. Time to Nail Breakage by Nail Brand

	Time to Nail Breakage	
Nail Brand	Mean (Months)	Median (Months)
TFNA	XX.XX	XX.XX
Gamma3	XX.XX	XX.XX
Natural Nail	XX.XX	XX.XX
All Nails	XX.XX	XX.XX

5.3 Secondary Objective 3

The third secondary objective is to estimate and compare the incidence rate of nail breakage by diameter of cephalomedullary nail (~15mm and ~17mm). DePuy Synthes TFNA, Stryker Gamma3, and Zimmer Natural Nail have nails with diameter ~15mm, which will be pooled together. DePuy Synthes TFN nail has a diameter of ~17mm. Incidence rates will be reported by nail diameter for the first 18 months. See Table Shell 9.

Table Shell 9. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Nail Diameter

	Time
Nail Diameter	0-18 months
~15mm	XX.XX(XX.XX,XX.XX)
~17mm	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

5.4 Secondary Objective 4

The fourth secondary objective is to estimate nail breakage incidence rates by calendar year stratified by diameter of cephalomedullary nail (~15mm or ~17mm). DePuy Synthes TFNA, Stryker Gamma3, and Zimmer Natural Nail have nails with diameter ~15mm, which will be pooled together. DePuy Synthes TFN nail has a diameter of ~17mm. Incidence rates will be reported by nail diameter and calendar year of the index procedure for the first 18 months. See Table Shell 10.

Table Shell 10. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Nail Diameter and Calendar Year for the First 18 months.

	Nail Diameter	
Calendar Year	~15mm	~17mm
2010	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2011	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2012	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2013	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2014	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2015	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2016	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2017	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2018	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)
2019	XX.XX(XX.XX,XX.XX)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

5.5 Secondary Objective 5

The fifth secondary objective is to estimate nail breakage incidence rates stratified by length of cephalomedullary nail (short or long). Short nails are defined as ≤ 235 mm and long nails > 235 mm. Incidence rates will be reported by nail length for the first 18 months, See Table Shell 11.

Table Shell 11. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Nail Length

	Time
Nail Length	0-18 months
Short (≤ 235)	XX.XX(XX.XX,XX.XX)
Long (> 235 mm)	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

5.6 Secondary Objective 6

The sixth secondary objective is to estimate nail breakage incidence rates stratified by patient characteristics including: age, sex, race, type of fracture, obesity diagnosis, presence of pathologic fracture, and presence of polytrauma. Incidence rates will be reported by patient characteristics for the first 18 months. See Table Shell 12.

Table Shell 12. Incidence Rates Per 1000 Patient-Months and Interval Estimates (95%) by Patient Characteristics

	Time

http://dx.doi.org/10.2106/JBJS.20.02002

Patient Characteristic		0-18 months
Age		
	<65	XX.XX(XX.XX,XX.XX)
	≥65	XX.XX(XX.XX,XX.XX)
Sex		
	Male	XX.XX(XX.XX,XX.XX)
	Female	XX.XX(XX.XX,XX.XX)
Race		
	Black	
	White	
	Other	
Fracture Type		
	Subtrochanteric	XX.XX(XX.XX,XX.XX)
	Pertrochanteric	XX.XX(XX.XX,XX.XX)
	Intercapsular	XX.XX(XX.XX,XX.XX)
	Femoral Shaft	XX.XX(XX.XX,XX.XX)
	All Others	XX.XX(XX.XX,XX.XX)
Obesity		
	Yes	XX.XX(XX.XX,XX.XX)
	No	XX.XX(XX.XX,XX.XX)
Pathological Fracture		
	No pathologic fracture	XX.XX(XX.XX,XX.XX)
	Pathologic fracture	XX.XX(XX.XX,XX.XX)
Polytrauma		
	No Polytrauma	XX.XX(XX.XX,XX.XX)
	Polytrauma	XX.XX(XX.XX,XX.XX)

Note: Variance estimation uses cluster robust standard errors with a finite cluster bias correction of C/(C-1) and no correction for the heteroscedasticity-consistent covariance estimate (i.e., HC0).

References

1. Williams RL. Product-Limit survival functions with correlated survival times. *Lifetime Data Analysis*. 1995; 1:171-186.
2. Lumley T. Survival curve estimation under complex sampling [Internet]. 2020 [cited 2020 May 30]; Available from: http://r-survey.r-forge.r-project.org/survey/survcurve.pdf
3. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005; 24:1713–1723.
4. Austin PC, Wagner P, Merlo J. The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis. *Statistics in Medicine*. 2017; 36: 928-938.
5. Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. *Biometrics* 2000; 56:909–914.
6. Rosenbaum PR. *Observational Studies*. Springer: New York, NY, 2009.

7. Rubin, D.B. The design versus the analysis of observational studies of causal effects: parallels with the design of randomized trials. *Statistics in Medicine* 2007; 26: 20-36.
8. Cafri G, Paxton EW. Mitigating reporting bias in observational studies using covariate balancing methods. *Observational Studies* 2018*;* 4: 292-296.
9. Yue LQ. Regulatory considerations in the design of comparative observational studies using propensity scores. *Journal of Biopharmaceutical Statistics* 2012; 22:1272–1279.
10. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 2004; 99:609–618
11. Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. *Statistics in Medicine* 2015; 34:3949-3967.
12. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One* 2011; 6:e18174.
13. Austin PC and Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical Methods in Medical Research* 2016; 25: 2214–2237.
14. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; 23:2937–2960.
15. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010; 29:2137–2148.
16. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; 46: 399-424
17. Yang D, Dalton DE. A unified approach to measuring the effect size between two groups using SAS. SAS Global Forum 2012, Paper 335-2012.
18. Finner H, Strassburger K. The partitioning principle: a powerful tool in multiple decision theory. *The Annals of Statistics* 2002; 30:1194--1213.
19. Goeman JJ, Solari A, Stijnen, T. Three-sided hypothesis testing: simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine 2010;* 29: 2117–2125.
20. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 2005; 24:3089-3110.
21. Cafri G, Lesko J. Clustered survival data in multi-site single-arm clinical studies**.** Poster presented at: AdvaMed Virtual Medical Devices and Diagnostics Statistical Issues Poster Session; April 23, 2020; Virtual.
22. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 2013; 32: 2837–2849.
23. Lee EW, Wei LJ and Amato D. Cox-type regression analysis for large number of small groups of correlated failure time observations. In: Klein JP and Prem Goel P (eds) Survival analysis, state of the art. Netherlands: Kluwer, 1992, pp.237–247.

http://dx.doi.org/10.2106/JBJS.20.02002

24. Cafri G, Wang W, Chan P, Austin PC. A review of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research* 2019*;* 28: 3142-3162.
25. Link CL. Confidence intervals for the survival function using Cox's proportional hazards model with covariates. *Biometrics* 1984; 40: 601-610.
26. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;21:383-388.