Supplement to Vail EA, Feng R, Sieber F, et al.

Long-term outcomes with spinal versus general anesthesia for hip fracture surgery

Statistical Analysis Plan

Original Analysis Plan (version date: May 5, 2022) .....Page 2
Final Analysis Plan (version date: December 2, 2022).....Page 7
Table 1. Summary of changes to statistical analysis plan.....Page 12

REGAIN Statistical Analysis Plan for Long-term Patient Reported Outcomes

Quality Control Prior to Unblinding

The REGAIN randomized trial was conducted with blinding to treatment assignment of all study staff involved in data collection for the primary outcome and its components and blinding of study investigators to subject randomization assignment in all assessments of aggregate data related to patient characteristics and study outcomes. Prior to unblinding, we will review the distributions of each variable to be included in the analysis, aggregated across treatment arms, to identify any outlying values that may be inaccurate and should be checked. To the extent possible, inaccuracies will be resolved and the database updated with the correct values. Data that are clearly incorrect but cannot be corrected will be excluded from the analyses. Data that are unusual but not impossible, and cannot either be verified or corrected, will remain in the analysis.

Baseline Data

Baseline (pre-randomization) demographic and clinical variables will be examined to evaluate general trends and determine whether there are any notable imbalances that may lead to secondary adjustments. Continuous variables will be summarized through standard measures of central tendency and spread including means, medians, standard deviations and interquartile ranges (IQRs). Frequency distributions will be calculated for categorical variables.

General Principles

Intention-to-treat principle and primary analysis sample: Analyses will follow the intention-to-treat principle, with sensitivity analyses performed to assess the potential impact of missing values and noncompliance. Subjects who were assigned randomization codes due to purely technical errors in operating the database will be excluded from any analysis. Subjects who died before receiving either study treatment will be also excluded from any analysis. Subjects randomized and later found not to meet eligibility criteria will be included in the primary analysis to avoid potential bias due to differential assessment of eligibility. Sensitivity analyses excluding these subjects will be conducted.

Adjustment for stratification factors and country of randomization: All analyses (primary, secondary, exploratory) will adjust for sex, fracture type (femoral neck vs intertrochanteric/subtrochanteric), and country of randomization (U.S., Canada). Sex, facture type, and site were the original stratification factors for randomization assignment, but with 46 participating sites, some with very few enrolled subjects, it will not be possible to include site in the stratified models. Because there might be systematic differences in health care and insurance use between Canada and U.S. sites, we will add country as an additional stratification factor. Subjects who were randomized into the wrong stratum will be analyzed as per-randomization strata.

Principles of statistical hypothesis testing and p-value reporting: All analyses will use 2-sided tests and an overall significance level of 0.05 as the threshold for statistical significance. We will report nominal p-values.

Outcome Specifications

Primary long-term outcome: *Days from randomization to death*, censored at post-randomization day 365, will be analyzed as a time-to-event outcome. Survival status and date of death will be ascertained

based on telephone interviews with participants or appropriate proxy informants or the site. A National Death Index (NDI) search for mortality information will be done for individuals with unknown vital status at one year. For subjects with partial date-of-death data (i.e. month and year only), the date of death will be imputed as the 15th day of the month. Survival time for individuals with unknown vital status at one year and no death report in the NDI will be censored at the date last known to be alive.
Survival outcomes will be compared between arms using a Cox proportional hazards regression model, adjusted for sex, fracture type, and country. The proportional hazards assumption will be assessed using both failure-time graphs and formal statistical tests of zero slopes in the Schoenfeld residuals. If the proportional hazards assumption is violated, advanced techniques such as including a time-dependent covariate in the model will be applied (Grambsch and Therneau, 1994; Hess, 1995). Kaplan-Meier curves displaying failure times by study arm for each outcome will be presented.

Secondary outcomes: We will evaluate two secondary outcomes: death or debility and ambulation.

- *Death or debility (move from independent living to a care facility) at one year.* We will identify the location of residence on or around post-randomization days 60, 180, and 365 via blinded telephone interview with the patient or proxy. The analysis for this endpoint will include only individuals who were living independently (at their own home or a retirement home) at the time of study entry, and whose residence or death status at day 365 was known. Individuals who had died by post-randomization day 365 or had moved to a nursing home as reported on the day 365 survey will be considered positive for debility. Those still living independently (at home or an independent living retirement home) as reported on the day 365 survey will be considered negative for debility. Debility will be compared between the two arms, using the Mantel-Haenszel (MH) test for difference in proportions, stratified by sex, fracture type, and country, and the relative risk (RR) of debility will be calculated. The Breslow-Day test (Breslow and Day, 1980) will be used to assess homogeneity of the RRs across strata. If there is statistically significant heterogeneity ($p<0.05$), we will test separately and report separate RRs within strata.

  If a substantial amount of data on residence at one year is missing, we will analyze this outcome using a multiple imputation approach, including baseline and outcome variables at days 60 and 180 in the model.

- *Ability to walk 10 feet without human assistance at post-randomization days 60, 180, and 365*. . This endpoint will be analyzed as a longitudinal binary outcome among 365 post-randomization day survivors with any informative ambulation assessment during three follow-up windows. Ambulatory status was assessed via blinded telephone interview with the subjects or their proxies. The 60-day survey was conducted within a window of assessment beginning on post-randomization day 30 and ending on post-randomization day 90; the window for the 180-day survey was from post-randomization day 135 to day 225; the window for the 365-day survey was from post-randomization day 305 to day 425. Individuals who have died by post-randomization day 365 will be excluded from the analysis.

  Subjects for whom ambulation information is missing on all three surveys (e.g., due to loss to follow-up or subject refusal) will be excluded from the analysis. Sensitivity analyses, described below, will be conducted to assess the potential influence of missing data on the study results. The treatment effect will be assessed by logistic mixed effects regression model (GLMM), adjusted for sex, fracture type, country, and days since randomization at assessment. A random

intercept term per individual with an unstructured variance-covariance matrix will be included to allow for within-subject dependence.

In a secondary analysis we will incorporate an interaction term between treatment and a function of time in the GLMM to assess the possibility of variability of treatment effect over time. Both main and interaction effects of the treatment will be reported and tested for significance.

- *Death or inability to ambulate at one year.* The analysis for this endpoint will include all individuals with death or walking assessment at one year. Individuals who had died by post-randomization day 365 or were unable to walk 10 feet without human assistance as reported on day 365 survey, will be considered positive for this composite outcome. The remaining individuals who had a walking assessment on the day 365 survey will be considered negative for this outcome. The outcome will be compared between the two arms, using the Mantel-Haenszel (MH) test for difference in proportions, stratified by sex, fracture type, and country, and the relative risk (RR) of debility will be calculated. The Breslow-Day test (Breslow and Day, 1980) will be used to assess homogeneity of the RRs across strata. If there is statistically significant heterogeneity ($p<0.05$), we will test separately and report separate RRs within strata.

Analyses for Heterogeneity of Treatment Effects

We will evaluate treatment difference in subgroups when tests of interaction indicate that the treatment effect differs across levels of the subgroups. The following variables define subgroups that will be evaluated for heterogeneity of treatment effects:

- Sex
- Fracture type: Actual femoral neck vs intertrochanteric/subtrochanteric
- Country of enrollment: Canada vs US
- Need for assistive devices to ambulate prior to fracture
- Age category: 85 or older, under 85
- Community residence prior to fracture: independent living (home or retirement home) vs nursing home, rehabilitation or acute care hospital vs other.

The first step will be to test for interaction between treatment and the variable defining the subgroups in a Cox model (survival outcome) or logistic model (binary outcome). To reduce small-sample bias, we will use Firth's penalized logistic regression method (Firth, 1993; King and Zeng, 2001) when the event rate of an outcome is less than 5%. Analysis within subgroups will be performed for any variable showing an interaction yielding a p-value of 0.20 or less. Subgroup analyses will utilize the same methods as the primary analysis.

Sensitivity Analysis for Missing Ambulation Data

The primary analysis will use complete data, excluding individuals without any outcome data. We will assess the potential for bias incurred by ignoring the missing data using several approaches. First, baseline characteristics of patients with missing ambulation outcome at all 3 time points will be compared with the baseline characteristics of those included in the analysis to assess potential for bias incurred by excluding the patients with missing data. To further assess the potential for bias incurred by

ignoring patients with missing data, we will conduct an analysis under the assumption of data missing at random using the inverse-probability-weighting method (Hogan and Lancaster, 2004). In the GLMM, each subject at each time point will be weighted by the inverse probability of being a complete case at that time. The probability at a particular time point will be estimated from a logistic regression on missingness using sex, fracture type, country, treatment, and previous outcome(s) before that time point.

To avoid potential survivor bias, we will consider a joint model for both survival and longitudinal categorical outcomes (Choi et al., 2015).

Sensitivity Analysis for Noncompliance

To assess the potential impact of non-compliance on the study outcomes, we will use an instrumental variable approach to estimate the average causal effect of spinal anesthesia vs general anesthesia on survival outcome. Randomization assignment will serve as the instrumental variable because it is designed to be independent of unobserved confounders.

Subjects who received both spinal and general anesthesia will be categorized as having received general anesthesia. We will use a structural proportional hazard model to estimate the causal hazard ratio, adjusted for sex, fracture type, and country (Loeys et al., 2005).

References:

1. Breslow NE, Day NE. 1980. Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.
2. Hogan WH and Lancaster T. 2004. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research, 13:17–48.
3. Raghunathan TW, Lepkowksi JM, Van Hoewyk J, Solenbeger P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27:85–95.
4. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. 2007. Statistical Methods in Medical Research,16:219–242.
5. Sitlani CM, Heagerty PJ, Blood EA, Tosteson TD. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. Stat Med. 2012 Jul 20;31(16):1738-60.
6. Bond SJ, White IR, Sarah Walker A. Instrumental variables and interactions in the causal analysis of a complex clinical trial. Stat Med. 2007 Mar 30;26(7):1473-96.
7. Greenland S. 2000. An introduction to instrumental variables for epidemiologists, International Journal of Epidemiology. 2000; 29(4): 722–729.
8. Loeys, T., Goetghebeur, E. & Vandebosch, A. Causal Proportional Hazards Models and Time-constant Exposure in Randomized Clinical Trials. *Lifetime Data Anal* **11,** 435–449 (2005).
9. Grambsch PM, Therneau TM. 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81: 515–526.
10. Hess KR. 1995. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine, 14: 1707–1723.
11. Choi J, Cai J, Zeng D, Olshan AF. Joint Analysis of Survival Time and Longitudinal Categorical Outcomes. *Stat Biosci*. 2015;7(1):19-47.

REGAIN Statistical Analysis Plan for Long-term Patient Reported Outcomes

Quality Control Prior to Unblinding

The REGAIN randomized trial was conducted with blinding to treatment assignment of all study staff involved in data collection for the primary outcome and its components and blinding of study investigators to subject randomization assignment in all assessments of aggregate data related to patient characteristics and study outcomes. Prior to unblinding, we will review the distributions of each variable to be included in the analysis, aggregated across treatment arms, to identify any outlying values that may be inaccurate and should be checked. To the extent possible, inaccuracies will be resolved and the database updated with the correct values. Data that are clearly incorrect but cannot be corrected will be excluded from the analyses. Data that are unusual but not impossible, and cannot either be verified or corrected, will remain in the analysis.

Baseline Data

Baseline (pre-randomization) demographic and clinical variables will be examined to evaluate general trends and determine whether there are any notable imbalances that may lead to secondary adjustments. Continuous variables will be summarized through standard measures of central tendency and spread including means, medians, standard deviations and interquartile ranges (IQRs). Frequency distributions will be calculated for categorical variables.

General Principles

Intention-to-treat principle and primary analysis sample: Analyses will follow the intention-to-treat principle, with sensitivity analyses performed to assess the potential impact of missing values and noncompliance. Subjects who were assigned randomization codes due to purely technical errors in operating the database will be excluded from any analysis. Subjects who died before receiving either study treatment will be also excluded from any analysis. Subjects randomized and later found not to meet eligibility criteria will be included in the primary analysis to avoid potential bias due to differential assessment of eligibility. Sensitivity analyses excluding these subjects will be conducted.

Adjustment for stratification factors and country of randomization: All analyses (primary, secondary, exploratory) will adjust for sex, fracture type (femoral neck vs intertrochanteric/subtrochanteric), and country of randomization (U.S., Canada). Sex, facture type, and site were the original stratification factors for randomization assignment, but with 46 participating sites, some with very few enrolled subjects, it will not be possible to include site in the stratified models. Because there might be systematic differences in health care and insurance use between Canada and U.S. sites, we will add country as an additional stratification factor. Subjects who were randomized into the wrong stratum will be analyzed as per-randomization strata.

Principles of statistical hypothesis testing and p-value reporting: All analyses will use 2-sided tests and an overall significance level of 0.05 as the threshold for statistical significance. We will report nominal p-values.

Outcome Specifications

Primary long-term outcome: *Days from randomization to death*, censored at post-randomization day 365, will be analyzed as a time-to-event outcome. Survival status and date of death will be ascertained

based on telephone interviews with participants or appropriate proxy informants or the site. A National Death Index (NDI) search for mortality information will be done for individuals with unknown vital status at one year. For subjects with partial date-of-death data (i.e. month and/or year only), the date of death will be imputed as the 15th day of the month, or at the midpoint between the last date known alive and the date that the subject was reported to be deceased. Survival time for individuals with unknown vital status at one year and no death report in the NDI will be censored at the date last known to be alive.

Survival outcomes will be compared between arms using a Cox proportional hazards regression model, adjusted for sex, fracture type, and country. The proportional hazards assumption will be assessed using both failure-time graphs and formal statistical tests of zero slopes in the Schoenfeld residuals. If the proportional hazards assumption is violated, advanced techniques such as including a time-dependent covariate in the model will be applied (Grambsch and Therneau, 1994; Hess, 1995). Kaplan-Meier curves displaying failure times by study arm for each outcome will be presented.

Secondary outcomes:

- *Death or debility (move from independent living to a care facility) at one year.* We will identify the location of residence on or around post-randomization days 60, 180, and 365 via blinded telephone interview with the patient or proxy. The analysis for this endpoint will include only individuals who were living independently (at their own home or a retirement home) at the time of study entry, and whose residence or death status at day 365 was known. Individuals who had died by post-randomization day 365 or had moved to a nursing home as reported on the day 365 survey will be considered positive for debility. Those still living independently (at home or an independent living retirement home) as reported on the day 365 survey will be considered negative for debility. Debility will be compared between the two arms, using the Mantel-Haenszel (MH) test for difference in proportions, stratified by sex, fracture type, and country, and the odds ratio (OR) of debility will be calculated. The Breslow-Day test (Breslow and Day, 1980) will be used to assess homogeneity of the ORs across strata. If there is statistically significant heterogeneity ($p<0.05$), we will test separately and report separate ORs within strata. If a substantial amount of data on residence at one year is missing, we will analyze this outcome using a multiple imputation approach, including baseline and outcome variables at days 60 and 180 in the model.

- *Ability to walk 10 feet without human assistance at post-randomization days 60, 180, and 365*. This endpoint will be analyzed as a longitudinal binary outcome among 365 post-randomization day survivors with any informative ambulation assessment during three follow-up windows. Ambulatory status was assessed via blinded telephone interview with the subjects or their proxies. The 60-day survey was conducted within a window of assessment beginning on post-randomization day 30 and ending on post-randomization day 90; the window for the 180-day survey was from post-randomization day 135 to day 225; the window for the 365-day survey was from post-randomization day 305 to day 425. Individuals who have died by post-randomization day 365 will be excluded from the analysis.

  Subjects for whom ambulation information is missing on all three surveys (e.g., due to loss to follow-up or subject refusal) will be excluded from the analysis. Sensitivity analyses, described below, will be conducted to assess the potential influence of missing data on the study results.

The treatment effect will be assessed by logistic mixed effects regression model (GLMM), adjusted for sex, fracture type, country, and days since randomization at assessment. A random intercept term per individual with an unstructured variance-covariance matrix will be included to allow for within-subject dependence.

In a secondary analysis we will incorporate an interaction term between treatment and a function of time in the GLMM to assess the possibility of variability of treatment effect over time. Both main and interaction effects of the treatment will be reported and tested for significance.

- *Death or inability to ambulate at one year.* The analysis for this endpoint will include all individuals with death or walking assessment at one year. Individuals who had died by post-randomization day 365 or were unable to walk 10 feet without human assistance as reported on day 365 survey, will be considered positive for this composite outcome. The remaining individuals who had a walking assessment on the day 365 survey will be considered negative for this outcome. The outcome will be compared between the two arms, using the Mantel-Haenszel (MH) test for difference in proportions, stratified by sex, fracture type, and country, and the odds ratio (OR) of debility will be calculated. The Breslow-Day test (Breslow and Day, 1980) will be used to assess homogeneity of the ORs across strata. If there is statistically significant heterogeneity ($p<0.05$), we will test separately and report separate ORs within strata.

Analyses for Heterogeneity of Treatment Effects

We will evaluate treatment difference in subgroups when tests of interaction indicate that the treatment effect differs across levels of the subgroups. The following variables define subgroups that will be evaluated for heterogeneity of treatment effects:

- Sex
- Fracture type: Actual femoral neck vs intertrochanteric/subtrochanteric
- Country of enrollment: Canada vs US
- Need for assistive devices to ambulate prior to fracture
- Age category: 85 or older, under 85
- Community residence prior to fracture: independent living (home or retirement home) vs nursing home, rehabilitation or acute care hospital vs other.

The first step will be to test for interaction between treatment and the variable defining the subgroups in a Cox model (survival outcome) or logistic model (binary outcome). To reduce small-sample bias, we will use Firth's penalized logistic regression method (Firth, 1993; King and Zeng, 2001) when the event rate of an outcome is less than 5%. Analysis within subgroups will be performed for any variable showing an interaction yielding a p-value of 0.20 or less. Subgroup analyses will utilize the same methods as the primary analysis.

Sensitivity Analysis for Missing Ambulation Data

The primary analysis will use complete data, excluding individuals without any outcome data. We will assess the potential for bias incurred by ignoring the missing data using several approaches. First, baseline characteristics will be compared between patients with and without complete 365-day follow-up survival data to assess potential for bias incurred by incomplete data. , we will conduct a sensitivity

analysis used an imputation method for the Cox model, which allows for possible informative censoring (Jackson et al., 2014). For patients without complete data, new event time and censoring time will be imputed. If there is any difference in the baseline characteristics between the patients with and without 365-day complete data, we will use subjects' baseline characteristics to determine subject-specific hazard change at the point of censoring; if there is no difference in the baseline characteristics between two groups of patients, the same hazard will be assumed before and after the censoring time. We will impute samples for 1,000 times.

Sensitivity Analysis for Noncompliance

To assess the potential impact of non-compliance on the study outcomes, we will use an instrumental variable approach to estimate the average causal effect of spinal anesthesia vs general anesthesia on survival outcome. Randomization assignment will serve as the instrumental variable because it is designed to be independent of unobserved confounders.

Subjects who received both spinal and general anesthesia will be categorized as having received general anesthesia. We will use a structural proportional hazard model to estimate the causal hazard ratio, adjusted for sex, fracture type, and country (Martinussen et al., 2019).

References:

1. Breslow NE, Day NE. 1980. Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.
2. Hogan WH and Lancaster T. 2004. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research, 13:17–48.
3. Raghunathan TW, Lepkowksi JM, Van Hoewyk J, Solenbeger P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27:85–95.
4. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. 2007. Statistical Methods in Medical Research,16:219–242.
5. Sitlani CM, Heagerty PJ, Blood EA, Tosteson TD. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. Stat Med. 2012 Jul 20;31(16):1738-60.
6. Bond SJ, White IR, Sarah Walker A. Instrumental variables and interactions in the causal analysis of a complex clinical trial. Stat Med. 2007 Mar 30;26(7):1473-96.
7. Greenland S. 2000. An introduction to instrumental variables for epidemiologists, International Journal of Epidemiology. 2000; 29(4): 722–729.
8. Grambsch PM, Therneau TM. 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81: 515–526.
9. Hess KR. 1995. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine, 14: 1707–1723.
10. Jackson D, White I, Seaman S, Evans H, Baisley K, Carpenter J. 2014. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. Statistics in Medicine, 33(27):4681–4694.Martinussen T, Sørensen DN, Vansteelandt S. 2019. Instrumental variables estimation under a structural Cox model, Biostatistics, 20(1): 65–79.

Original SAP page number(s)	Change made	Date of change
2, 3	Modified to specify the calculation of odds ratios (versus risk ratios) for two outcomes: (1) Death or debility at one year (2) Death or inability to ambulate at one year	October 26, 2022
3-4	Modified to add a comparison of baseline characteristics of patients with and without complete 365-day survival data.	December 2, 2022
4	Modified to specify multiple imputation as the sensitivity analysis method to assess potential impact of informative censoring in survival analysis.	December 2, 2022

SAP: statistical analysis plan

Original statistical analysis plan was completed May 5, 2022. Analyses were completed January 3, 2022.

Table 1. Summary of changes to statistical analysis plan