Final Statistical Analysis Plan

(according to the paper submitted to the JAMA Neurology and its supplementary material)

Abbreviations:

ACDF	Anterior Cervical Discectomy with Fusion
EQ-5D-5L	EuroQol 5-Dimensions 5-Level
FACET	Foraminotomy ACDF Cost-Effectiveness Trial
FOR	Posterior Cervical Foraminotomy
NDI	Neck Disability Index
SAP	Statistical Analysis Plan
VAS	Visual Analogue Scale
WAS	Work Ability Index Scale

Research aims and hypotheses:

Primary hypotheses:

- The effectiveness of the FOR technique is noninferior compared with the ACDF technique
- The FOR technique is cost-effective compared with the ACDF technique (a cost-effectiveness analysis after 2-years will be conducted separately)

Secondary hypotheses:

- The FOR technique will have a lower complication rate compared with the ACDF technique.
- The FOR technique will have lower direct and indirect costs compared with the ACDF technique.
- The FOR technique is associated with more neck pain in the first 30 days after the surgical procedure.

Study design and treatment allocation:

The FACET study is a prospective, multicenter, investigator-blinded randomized controlled trial with a follow-up of 2 years. Both the FOR (experimental group) and the ACDF (active control) are established surgical techniques. The randomization was performed using an independent institute web-based block randomization design, stratified by center. Blinding of the participant or the surgeon is not feasible. The data analysis will be performed with blinded data. Assessment of Odom criteria will be performed by an independent interviewer who is blinded to the treatment allocation.

Non-inferiority margin

A non-inferiority trial design is chosen to show whether FOR (experimental) has at least as much efficacy as ACDF (active control) or is worse by an amount less than 10% with regards to a successful outcome (Odom score) and decrease in arm pain (Visual Analogue Scale (VAS)) after surgery. No evidence-based non-inferiority margin for this research question exists. Therefore, based on the assumption that FOR reduces costs, has fewer complications, leads to less work absenteeism, and leads to higher quality of life, the non-inferiority margin of 10% was chosen, that would justify a tolerable loss of efficacy.

Sample size

A sample size of 308 patients was calculated that would give the trial 80% power to rule out a between-group difference in the success rate with an alpha of 0.05, and a drop-out ratio of 10%, and non-inferiority margin of 10%. During the trial, a lower inclusion rate was observed and several measures to improve inclusion were taken. This included frequent newsletters, defined instructions for surgeons to include participants and the enrollment of additional medical centers. Nonetheless, a delay in inclusion could not be fully alleviated. When also the COVID-pandemic started, with an

extreme reduction of non-emergent health care, an interim analysis and post-hoc power calculation were performed by a statistician who was not involved in the study design.

The interim-analysis included participants that completed the 1-year follow-up and those that completed the 2-year follow-up at that time (the 2-year follow-up is still ongoing, thus not completed for all patients yet). The analysis included the primary outcome on which the sample size was based on; a successful Odom score ('excellent' or 'good'). Besides a complete case analysis, additional sensitivity analyses were performed; a scenario with full cases in which all missing cases were considered to have an unsuccessful outcome; a scenario with predefined sample sizes (140 participants per group) with all remaining missing data coded as unsuccessful; and a scenario in which additional participants in the ACDF group were coded as unsuccessful (stacking the Odds against FOR) until the noninferiority margin was reached (Figure 1). All scenario's indicated non-inferiority of FOR with a confidence interval within the predefined delta of 0.1. Therefore, we could conclude that it was safe to end the inclusion at 86% of the predefined sample size.

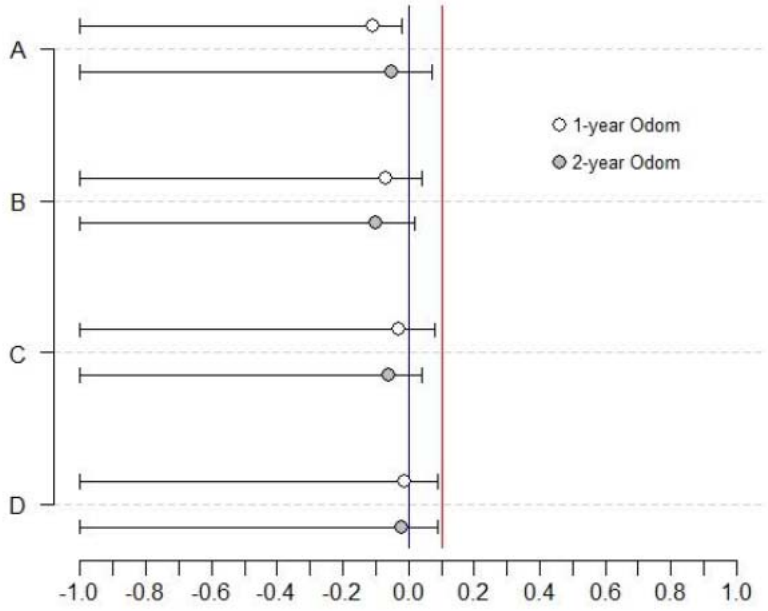

Figure 1. Interim-analyses of non-inferiority.

Point estimate + one-sided 95% confidence interval (CI) are depicted. Different scenarios were calculated for the proportion of a successful primary outcome (Odom score): 86 versus 81 in the FOR and ACDF group after 1-year of follow-up, and 69 versus 60 after 2-years of follow-up, respectively. **the 2-year follow-up is still ongoing and thus not completed for all patients. The red line depicts the non-inferiority margin of 10%.

(A) Intention-to-treat analysis with complete cases after 1- and 2-years of follow-up. **(B)** Intention-to-treat analysis with full cases (FOR 119 and ACDF 124), with missing data coded as unsuccessful. **(C)** Analysis with predefined sample sizes (140 patients per group) with missing data coded as unsuccessful **(D)** Scenario where additional patients in the ACDF group were coded as having a successful outcome, until the non-inferiority margin was reached (stacking the odds against FOR).

Post-hoc power analysis

In order to estimate the achieved power, post-hoc power analyses were conducted using the complete cases and full data for both primary outcomes. The number of patients with a successful Odom scorewas identical in the two analyses (86 in the FOR group and 81 in the ACDF group), and as expected the total number of patients was smaller in the complete case than the full data analysis (98 versus 119 in the PCF group and 106 versus 124 in the ACDF group). Using these sample sizes and the actual achieved proportions of the successful Odom score, two-proportion z-tests with a one-sided 0.05 level of significance were performed. With a noninferiority margin of 10%, the Odom score achieved a power of 0.98 in the complete case analysis and 0.85 in the full data analysis.

For the VAS arm pain at 1-year, there were 195 observations (96 in the FOR group and 99 in the ACDF group). With the abovementioned conditions, with an assumed standard deviation of 23, the power of the results was calculated to be 0.86. For the change score in arm pain at 1-year, 192 observations were collected (94 in the FOR group, 98 in the ACDF group). With an assumed standard deviation of 30, this led to a power of 0,63. All power calculations were performed with a Bonferroni correction, leading to an Alpha of 0.025.

Data collection methods*:

Preoperative history is obtained from standard care procedures and includes length, weight, number of months/years of neck and arm pain, signs and symptoms, other significant illnesses, pain medication (use of non-steroidal anti-inflammatory drugs) and smoking history. Information about the operative procedure will be obtained from the medical record of the participant and will include date and type of procedure, which level was operated, use of implants and occurrence of complications during the operative procedure. At baseline (i.e. at enrolment, before surgical procedure), participants will fill out web-based questionnaires. These questionnaires take ~30 min to fill in. The participant will visit the outpatient clinic 6 weeks after the surgical procedure, in line with standard care. Thereafter, at all follow-up moments including 1 and 2-years after follow-up, the participant fills out the same abovementioned questionnaires. An independent interviewer will contact the participants by telephone at all follow-up moments to assess the Odom criteria. During the complete period of the study, all adverse events will be reported. Adverse events are defined as undesirable experience occurring to a participant during the study, whether or not considered related to intervention. The definition of serious adverse events is in line with the guidelines of the International Council on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).[1]

Outcome measures:

The main primary outcome is the proportion of a successful score ('excellent' or 'good') on the modified Odom criteria (4-point rating scale) after 1 year of follow-up.[2]

The second primary outcome is a postoperative decrease in radiculopathy assessed by the VAS for self-reported arm pain (0-100; lower scores at 1-year and higher change scores from baseline indicating a greater decrease in pain).[3]

Secondary outcomes include differences in postoperative self-reported VAS-neck pain, quality of life, assessed with the EuroQol 5 Dimension 5 Level Survey (EQ-5D-5L)[4], neck disability assessed with the Neck Disability Index (NDI)[5], work ability, assessed with the Single-item Work Ability Index (WAI)[6], satisfaction with surgery, (serious) adverse events and reoperations.

Statistical methods: primary study parameters:

Noninferiority of the primary outcomes (a successful Odom score and the decrease of arm pain) after 1-year of follow-up will be assessed in the intention-to-treat population. Noninferiority will be tested with a two-proportion z-test at a one-sided 0.05 level of significance and noninferiority margin of 10%. Bonferroni corrections will be applied to adjust for multiple comparisons, leading to an alpha 0.025. Analyses will be conducted with and without continuity correction.

Detailed descriptive statistics will be provided for the data collected and one-sided 95% CIs will be calculated for all relevant estimates. Baseline characteristics will be calculated for the total randomized patient group and for the full cases. Also, the primary analysis will follow the per protocol principle. Sensitivity analysis are provided to evaluate robustness of the results.

Statistical methods: secondary study parameters:

Secondary end points will be analyzed in an exploratory manner after 1-year of follow-up at a two-sided 95% confidence interval , not adjusted for multiple comparison. T-tests for continuous data and z-tests for categorical outcomes will be used.

Additionally, a responder analysis of the secondary outcomes will be conducted, in which the proportion of patients with a response to treatment will be defined as those with an improvement from baseline up until 1 year after surgery, reaching or exceeding the predefined threshold of the prespecified mean clinical important difference (MCID).[7]

Missing data

Various post-hoc sensitivity analyses will be performed accounting for missing data. This will be done by simulating different scenarios and calculate the one-sided 95% CIs for: the full cases in which all missing cases are considered to have an unsuccessful outcome, and the predefined sample sizes in

both groups with all remaining data coded as unsuccessful. Additionally, as a measure of robustness, the fragility index will be calculated, which is the minimum number of patients whose status (primary outcome) has to change to convert a noninferior outcome to an inferior outcome.[8] This will be done by adding additional patients with a successful outcome to the ACDF group (thereby stacking the odds against FOR) until the noninferiority margin is reached.

References:

1. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH harmonized tripartite guideline: Clinical safety data management: Defenitions and standards for expedited reporting E2A*. Step 4 version ed. October.1994.http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficiacy/E2A/Step4/E2A_Guideline.pdf Accessed on 2nd of August, 2022.
2. Broekema AEH, Molenberg R, Kuijlen JMA, Groen RJM, Reneman MF, Soer R. The Odom Criteria: Validated at Last: A Clinimetric Evaluation in Cervical Spine Surgery. J Bone Joint Surg Am. 2019;101(14):1301–8.
3. Huskisson EC. Measurement of pain. *The Lancet*. 1974;304(7889). doi:10.1016/S0140-6736(74)90884-8
4. Soer R, Reneman MF, Speijer BLGN, Coppes MH, Vroomen PCAJ. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine J. 2012;12:1035-9.
5. Jorritsma W, De Vries GE, Dijkstra PU, Geertzen JHB, Reneman MF. Neck Pain and Disability Scale and Neck Disability Index: Validity of Dutch language versions. Eur Spine J. 2012;21:93-100.
6. De Zwart BCH, Frings-Dresen MHW, Van Duivenbooden JC. Test-retest reliability of the Work Ability Index questionnaire. Occup Med. 2002;52:177-81.
7. Broekema AEH, Kuijlen JMA, Lesman-Leegte GAT, et al. Study protocol for a randomised controlled multicentre study: The Foraminotomy ACDF Cost-Effectiveness Trial (FACET) in patients with cervical radiculopathy. BMJ Open. 2017;7(1):e012829.
8. Tignanelli CJ, Napolitano LM. The Fragility Index in Randomized Clinical Trials as a Means of Optimizing Patient Care. JAMA Surg. 2019;154(1):74–9.

Summary of changes in Statistical Analysis Plan

	Original protocol	Published protocol and present report	Summary of changes
Primary hypothesis	The effectiveness of the FOR technique is noninferior compared with the ACDF technique	The effectiveness of the FOR technique is noninferior compared with the ACDF technique	No changes
Study design and treatment allocation			
Noninferiority margin	A non-inferiority trial design is chosen to show whether FOR has at least as much efficacy as the ACDF technique or is worse by an amount <10% with regard to the primary outcome parameters.	A non-inferiority trial design is chosen to show whether FOR has at least as much efficacy as the ACDF technique or is worse by an amount <10% with regard to the primary outcome parameters.	No changes
Sample size	A sample size of 308 patients was calculated that would give the trial 80% power to rule out a between-group difference in the success rate with an alpha of 0.05, a drop-out ratio of 10%, and noninferiority margin of 10%.	The interim-analysis included participants that completed the 1-year follow-up and participants that included the 2-year follow-up at that time (the 2-year follow-up is still ongoing). The analysis was based on the prespecified primary outcome a successful Odom score ('excellent' or 'good'), on which the initial sample size was also based. Besides a 'complete case' analysis, additional sensitivity analyses were performed. All scenario's indicated non-inferiority of posterior cervical foraminotomy with a confidence interval within the predefined delta of 0.1. Therefore, we could conclude that it was safe	265 patients were included in the trial (86% of the predefined sample size) and 243 patients received the final treatment.

		to end the inclusion at 86% of the predefined sample size (265 patients).	
Outcome measures			
Primary outcome	The primary outcome is 'Operative success' defined as the postoperative decrease in radiculopathy assessed by the VAS for self-reported arm pain, between the patients operated with the FOR technique and with the ACDF technique during 24 months of follow-up. Additionally, 'Patient success' will be assessed by the modified Odom criteria, which address physical symptoms and socioeconomic status.	The main primary outcome is the proportion of a successful score ('excellent' or 'good') on the modified Odom criteria (4-point rating scale) after 1 year of follow-up.[2] The second primary outcome is a postoperative decrease in radiculopathy assessed by the VAS for self-reported arm pain (0-100; lower scores at 1-year and higher change scores from baseline indicating a greater decrease in pain).[3]	The successful Odom score after 1-year of follow-up ¥ is treated as our main primary outcome and the VAS-arm as our second primary outcome.**
Secondary outcomes	Secondary outcomes include differences in postoperative self-reported VAS-neck pain, quality of life, assessed with the EuroQol 5 Dimension 5 Level Survey (EQ-5D-5L), neck disability assessed with the Neck Disability Index (NDI), work ability, assessed with the Single-item Work Ability Index (WAI), satisfaction with surgery, (serious) adverse events and reoperations.	Secondary outcomes include self-reported VAS-neck pain, quality of life, assessed with the EuroQol 5 Dimension 5 Level Survey (EQ-5D-5L), neck disability assessed with the Neck Disability Index (NDI), work ability, assessed with the Single-item Work Ability Index (WAI), (serious) adverse events and reoperations.	No changes

Statistical analysis of primary outcome	Detailed descriptive statistics will be provided for the data collected and 95% CIs will be calculated for all relevant estimates. Clinical follow-up data will be analysed by analysis of covariance or generalised model alternatives for categorical or semiquantitative data. Changes within the treatment groups over time, as well as differences between groups, will be assessed by intention-to-treat analyses. Also, the primary analysis will follow the per protocol principle.	Detailed descriptive statistics will be provided for the data collected and **one-sided** 95% CIs will be calculated for all relevant estimates. Clinical follow-up data **after 1-year ¥** is analyzed using a two-proportion z-test at a one-sided 0.05 level of significance and noninferiority margin of 10%. Bonferroni correction will be applied to adjust for multiple comparisons. Changes within the treatment groups will be assessed by intention-to-treat analyses. Also, the primary analysis will follow the per protocol principle.	"One-sided" is added to the 95% CI. Besides the longitudinal analysis after 2 years, the clinical follow-up data will be analyzed after 1-year ¥ of follow-up with a two-proportion z-test (longitudinal analysis after 2 years is planned). ¶
Statistical analysis of secondary outcome	Secondary end points will be analysed in an exploratory manner at a two-sided significance level of 5%. Safety and tolerability parameters will be analysed descriptively. Analysis of time-dependent probabilities of critical events will be performed using the Kaplan-Meier method. Furthermore, multivariate event analyses will be performed using Cox proportional hazard regression models.	Secondary end points will be analyzed in an exploratory manner after 1-year of follow-up at a two-sided 90% confidence interval (equivalent to a one-sided 95% confidence interval). T-tests for continuous data and z-tests for categorical outcomes will be used. Additionally, a responder analysis of the secondary outcomes will be conducted.	t-test are used for continuous outcomes and z-tests for categorical outcomes. Also a responder analysis of the secondary outcomes is performed using the predefined MCIDs.
Missing data	For the purpose of a supportive sensitivity	Various post-hoc sensitivity analyses will	Various post-hoc sensitivity analyses are performed

	analysis, multiple imputation procedures will be applied.	be performed accounting for missing data.	regarding the missing data.
Other	British English is used in our original statistical analysis plan (according to the published protocol).	American English is used in our final statistical analysis plan.	Change of British English to American English.

****Outcome measures: primary outcomes:**

In the original SAP, the primary end point was defined as a successful Odom score as well as postoperative decrease of self-reported arm pain on the VAS-arm scale. It was stated in our original SAP that "*both items will be analyzed as appropriate depending on data distribution with a one-sided 0.05 level of significance (non-inferiority)". We consider the FOR technique non-inferior if the technique has at least as much efficacy (decreased VAS arm pain and the good/excellent scores on Odom criteria) as the ACDF technique or is worse by an amount <10%.*

In our final SAP we treat the successful Odom score as our main primary outcome and the VAS-arm as our second primary outcome.

We define the Odom criteria as main primary outcome because the sample size was calculated based on this outcome measure. Furthermore, the Odom criteria have been widely used in studies regarding different cervical procedures and are validated, assessing improvement of physical symptoms as well as the ability to perform daily activities (thus, the Odom is a more extended outcome measure than the VAS-arm score, assessing improvement of physical symptoms only).

For the second primary outcome, the decrease in arm pain, we chose to report both the arm pain score at 1 year and the change from baseline. This decision was made because in the original protocol the method of measuring the decrease in arm pain was not consistently defined. In our opinion, reporting both the 1 year result and the change score leads to optimal transparency of our results. To adjust for the multiple comparisons in the primary outcome, Bonferroni corrections were applied to all results.

¥ Analyses after 1-year of follow-up. Reason for change:

We had initially planned to report the results after all the participants completed the 2 years of follow-up. However, we decided to also report our results after 1-year of follow-up, because of the following reasons;

- Due to a lower inclusion rate than anticipated, partly due to the COVID-pandemic, the sample size of 308 patients was not reached and an interim-analysis was performed. The interim-analysis indicated a non-inferior outcome of FOR compared to ACDF at 86% of the

predefined sample size, with a low risk of type II error. Subsequently, robustness of this result was confirmed with different post-hoc sensitivity analyses. Because of the robustness of this result, and the existing controversy in spine surgery regarding treatment choice for patients with cervical radiculopathy (FOR or ACDF), we decided that our result as such would be of direct value for clinicians. As FOR and ACDF are widely used interventions in spine surgery, and the choice between the two is commonly encountered among surgeons and patients with cervical radiculopathy, the noninferiority of FOR compared to ACDF after one year can already be used to improve patient counselling and enhance shared decision making by surgeons and patients.

- The initial plan to do the longitudinal analysis after 2-years of follow-up is not changed.

¶ Statistical analysis of primary outcome. Reason for change:

We chose to present our results after 1-year using basic statistical analysis, to communicate our results in a prompt and straightforward fashion. As such, the results are good to interpret for clinicians and the robustness was clearly indicated with various sensitivity analyses. After 2-years of follow-up a longitudinal analysis and cost-effectiveness analysis will be performed.