

STATISTICAL ANALYSIS PLAN

for the TONOS study

Multicenter, national, randomized, patient-blinded, parallel group, non-inferiority trial to compare tonsillotomy versus extracapsular tonsillectomy in adults with obstructive sleep apnea and tonsil hypertrophy: The TONOS Trial.

ClinicalTrials.gov NCT07580170

Tommi Kauko and Jaakko Piitulainen

Version 4th November 2025

Table of Contents

INTRODUCTION3

STUDY DESIGN.....3

RESEARCH QUESTIONS AND HYPOTHESES4

STATISTICAL METHODS5

MODEL SPECIFICATIONS.....6

POWER AND SAMPLE SIZE CALCULATION8

INTERIM ANALYSES.....8

MISSING DATA AND SENSITIVITY ANALYSIS.....9

MULTIPLE TESTING AND ADJUSTMENT FOR MULTIPLICITY9

SOFTWARE9

REPORTING.....9

Introduction

This study aims to investigate whether tonsillotomy (TT) is non-inferior compared to extracapsular tonsillectomy (TE) in treating obstructive sleep apnea (OSA), by measuring home sleep study before and 4-6, 24 and 60 months after surgery.

Changes from baseline and after surgery are also recorded in validated questionnaires on daytime sleepiness (measured with Epworth Sleepiness Scale, ESS) and insomnia (measured with Insomnia Severity Index, ISI). Other secondary outcomes include quality of life (measured with 12-item General Health Questionnaire, GHQ-12), surgical success rate (according to Sher's criteria), surgical cure rate (defined as a total Apnea-Hypopnea Index (AHI) <5 after surgery), and decrease in nocturnal desaturation (measured with oxygen saturation SaO_2 , $\text{SpO}_2 < 90\%$, T90 and Oxygen Desaturation Index, ODI3).

The objective is to compare treatment effects between groups while accounting for longitudinal design and within-subject correlations.

Study Design

This is a multicenter, national, randomized, parallel-group, non-inferiority trial.

Randomization and Blinding:

Participants are blinded to the allocated intervention with an allocation ratio 1:1 using permuted blocks of four. Participants will be recruited from patients referred to an outpatient hospital clinic. The trial will be carried out in 4 tertiary hospitals in Finland.

Randomization sequence is generated by an independent statistician using SAS 9.4 and is implemented in REDCap on the day of or the morning before surgery.

The patients and healthcare personnel responsible for the treatment and care of the patient remain unaware of the used method of surgery until the 5-year follow-up is completed.

Variables To Be Collected:

- Patient demographics (sex, occupation, smoking (current and history), age, height, weight, body mass index (BMI))
- Baseline characteristics (tonsil size, tongue position, earlier and current treatment for OSA)
- Surgical findings (surgical instrumentation, surgical success rate, surgical cure rate)
- Primary outcomes: Apnea-Hypopnea Index (AHI) at 4-6, 24 and 60 months post-surgery
- Secondary outcomes:

- Daytime sleepiness (Epworth Sleepiness Scale, ESS), measured 6, 12, 24 and 60 months after surgery
- Insomnia severity (Insomnia Severity Index, ISI), measured 6, 12, 24 and 60 months after surgery
- Mental health (General Health Questionnaire, GHQ-12) and depression (Depression Scale, DEPS), measured 6, 12, 24 and 60 months after surgery
- Decrease in nocturnal desaturation in terms of oxygen saturation SaO₂ and SpO₂ and Oxygen Desaturation Index, ODI3

All analyses will be performed on the intention-to-treat (ITT) population, defined as all randomized participants who received the allocated intervention. A per-protocol (PP) analysis will also be reported for sensitivity, if necessary.

Research Questions and Hypotheses

Primary Research Question:

Does tonsillotomy result non-inferior compared to extracapsular tonsillectomy in apnea-hypopnea index at 4-6 months follow-up on patients with moderate or severe OSA and grade 2 to 4 tonsil size?

Secondary Research Questions:

Does tonsillotomy result non-inferior compared to extracapsular tonsillectomy in daytime sleepiness, insomnia symptoms, quality of life, surgical success rate, surgical cure rate and nocturnal desaturation.

Surgical success rate and surgical cure rate are related to AHI score.

Hypotheses:

Null: TT is inferior to TE, e.g. $TE - TT \geq \Delta$

Alternative: TT is non-inferior to TE, e.g. $TE - TT < \Delta$

where Δ is the non-inferiority margin. Above hypothesis is statistically tested by comparing the upper bound of the two-sided confidence interval for $TE - TT$ with the margin, which is specified in advance.

Statistical Methods

Descriptive Statistics:

Means and standard deviations are used to summarize patient demographics and baseline characteristics. Medians and interquartile ranges are used if variable has skewed distribution.

Frequencies and percentages are reported for categorical variables.

Comparability of Groups at Baseline:

- Continuous variables (age, height, weight, AHI): t-tests or Mann-Whitney U tests
- Categorical variables (sex, occupation, smoking status, tonsil size, tongue position, earlier treatment for OSA): chi-squared test or Fisher's exact test

Non-parametric equivalents are used when assumptions of the parametric tests are violated.

Primary Outcome Analysis:

Change from baseline to 4-6 months after surgery in Apnea-Hypopnea Index will be analysed using generalized linear model.

Independent variables to be included in the model are sex, age, smoking status, BMI, tonsil size and earlier or current treatment for OSA.

1 out of 4 of recruiting hospitals will record AHI at 24 and 60 months post-surgery. For this sub-population, separate generalized linear model with extended follow-up will be fitted with change from baseline to 24 months or 60 months after surgery in AHI or with repeated measures mixed model, if appropriate.

Secondary Outcomes Analysis:

Follow-up questionnaire at 12-months will be analysed with similar generalized linear model as the primary outcome.

Other assessments (ESS, ISI, GHQ-12, DEPS and NTSR) will be analysed with generalized linear mixed model to include random effects to the model. Multiple timepoints of the assessments are included as random effect to the model and within-subject correlations are accounted for using suitable covariance structure. The possible covariance structures are compound symmetry (CS), autoregressive (AR(1)) and Toeplitz. The best structure to be fitted is evaluated by Akaike's Information Criteria (AIC).

Link function is selected based on the dependent variable distribution.

Independent variables to be included in the model are sex, age, smoking status, BMI, tonsil size and earlier or current treatment for OSA.

Pairwise Comparisons:

Pairwise comparisons across time points will be adjusted using Tukey method for multiple comparisons. No adjustments will be applied over the tests and models of different outcomes.

Exploratory Analysis:

Associations between primary outcome variable and baseline assessments are explored by adding covariates to the mixed model.

Per-protocol analysis will be performed, if there is significant proportion of deviated participants from the intended randomized intervention.

Sensitivity analysis between recruiting hospitals will be performed by examining the differences in demographic and baseline variables.

Sensitivity analysis of subsets of subjects with and without prior treatments before surgery will be analysed with same generalized linear model as the primary outcome for all outcomes. Prior treatments that are investigated are CPAP (continuous positive airway pressure), MAD (mandibular advancement device) and positional therapy.

Change from baseline to 24 and 60 months after surgery in Apnea-Hypopnea Index will be analysed using generalized linear model.

Model Specifications

The generalized linear model (GLM) is specified as follows:

Outcome: Change in AHI

Fixed Effects:

- Treatment group
- Baseline AHI
- Sex
- Age
- Smoking status

- Tonsil size
- Body Mass Index
- Earlier or current treatment for OSA

Expressed as an equation:

$$Y = X\beta + \varepsilon$$

- Y: Change in AHI for subject i
- X: independent variables
- β 's: fixed effect coefficients
- $\varepsilon_i \sim N(0, \sigma^2)$: residual error

The generalized linear mixed model (GLMM) for repeated measures is specified as follows:

Outcome: <<assessment>>

Fixed Effects:

- Treatment group
- Baseline AHI
- Sex
- Age
- Smoking status
- Tonsil size
- Body Mass Index
- Earlier or current treatment for OSA

Random Effect:

- Patient ID to account for within-subject correlations.

Expressed as an equation:

$$g(\mu) = X\beta + Zb$$

- $g()$: link function
- μ : mean change in AHI
- X : independent variables
- β 's: fixed effect coefficients
- Z : random effects
- b 's: random effect coefficients

Secondary outcomes will be analyzed similarly using separate models for each outcome. For binary outcomes, the link function is logit whereas for continuous outcomes, identity link is used. Covariance matrix depends on the selected covariance structure.

Power and Sample Size Calculation

A power analysis was performed on August, 2025 for non-inferiority setup. Assuming a clinically relevant group difference to 10 events per hour in AHI (non-inferiority threshold) and with standard deviations derived from literature for TE and from a previous study for T2T, we calculated a pooled standard deviation estimate of 22.7 to be used in the calculations. Using the observed change in AHI from the aforementioned studies, the smaller difference in change revealed that 55 participants were needed in each group. Assuming 20% drop-out, the final sample size needed for this study is 132 participants total to achieve 80% statistical power in one-sided testing with 0.025 confidence level.

Interim Analyses

Two interim analyses will occur mid-recruitment. After 50 % and after 75 % of participants have completed the 4–6 month assessment.

Alpha-spending function: O'Brien–Fleming.

Decision rule: If the p-value for the primary endpoint is below the interim threshold, a recommendation to stop for efficacy will be issued. To control the Type I error, the thresholds of 0.0021 (50 % recruited) and 0.0097 (75 % recruited) are required to achieve to justify the early termination of the recruitment.

Stopping rules: No formal futility stopping; safety data (adverse events) will be reviewed at each interim.

Due the interim analyses, the P values in the statistical analyses less than 0.0215 will be considered statistically significant.

Missing Data and Sensitivity Analysis

Missing values are expected to be sparse (< 5% per variable per group). Complete case analysis will be used, and no imputation methods are applied.

Multiple Testing and Adjustment for Multiplicity

Pairwise comparisons between time points within each treatment group used the Tukey method. However, no adjustments were made for multiple testing across all outcomes (e.g., over primary and secondary outcome measures). This increases the risk of Type I error but allows for exploratory analysis without overly conservative thresholds.

Software

All analyses will be performed using R version 4.5.0 (or newer), with generalized linear models fitted using `lm()` function from stats package, mixed models fitted using `lmer()` function from the lme4 package and adjusted p-values calculated using the Tukey method via emmeans package. Figures will be produced using ggplot2 package.

Reporting

Results will be presented in accordance with CONSORT guidelines, including a flow diagram of participants through each phase (screening → randomisation → allocation → follow-up). Interim analysis reports will be included in a separate appendix.