PHACT

STATISTICAL ANALYSIS PLAN

VERSION 1.6

DATE: APRIL 8, 2024

Table of content

1 Revisions

Version	Date	Reason
1.0	May 25, 2017	Version 1.0 developed prior to enrolling first patient in the trial.
1.1	September 19, 2017	Minor clarifications after feedback from clinicians.
1.2	December 4, 2018	Updated sample size.
1.3	June 19, 2022	Finalization of SAP prior to database lock.
1.4	August 30, 2023	Specification of post-hoc analysis to compute relative risks for binary endpoints (Section 7.1).
1.5	January 19, 2024	Specification of post-hoc analysis of limb symmetry index (Section 7.2).
1.6	April 8, 2024	Specification of changes to the SAP as a result of the review process.

2 List of authors and reviewers

2.1 Authors

Martin Eklund, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

Kenneth Jonsson, Department of Surgical Sciences, Uppsala University, Sweden

Elsa Pihl, Department of Orthopedics Danderyd Hospital and Department of Clinical Science Danderyd Hopsital, Karolinska Institutet, Sweden

2.2 Reviewers

Frede Frihagen, Division of Orthopaedic Surgery, Oslo University Hospital, N-0450 Oslo, Norway

Olof Sköldenberg, Department of Orthopedics Danderyd Hospital and Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Sweden

Chiara Micoli, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

3 Introduction

The treatment of proximal hamstring avulsions is controversial. The literature suggests that operative treatment is superior to conservative, non-operative management. However, randomized evidence is lacking. Moreover, few of the existing observational studies have used validated outcome measures, such as Harris Hip Score, the Lower Extremity Functional Scale (LEFS), or the Perth hamstring assessment tool (PHAT).

The Proximal Hamstring Avulsion Clinical Trial (PHACT; NCT03311997) is a prospective, preference tolerant, multicentre, randomised, controlled non-inferiority trial with the aim to provide reliable evidence on how to treat physically active patients, 30–70 years of age, with proximal hamstring avulsions.

The planned analyses identified in this statistical analysis plan (SAP) will be included in future manuscripts. Exploratory analyses not necessarily identified in this SAP may be performed to support planned analyses. Any post-hoc exploratory or unplanned analyses not specified in this SAP will be identified as such in manuscripts for publication, and added as addenda to the SAP. The SAP may be updated during the course of the trial but will be finalized before database lock or any comparative analyses. Any further future analyses not specified in the analysis protocol will be documented in the revision history of this document.

4 Trial overview

4.1 Trial design

Patients with proximal hamstring avulsion will be randomly assigned to either operative treatment or to non-operative treatment. Both groups will follow the same standardized rehabilitation protocol.

Patients who are eligible for trial participations but where the patient or the treating physician equipoise to treatment cannot be reached, will be asked to participate in a parallel observational follow-up cohort with identical treatment options and follow-up. In the parallel cohort, the patient's/surgeon's preferred treatment is provided. Patients are followed for 24 months with study visits at 3, 6, 12, and 24 months.

An overview of the trial design is shown in Figure 1. The study design, interventions, eligibility criteria, and conduct are outlined in detail in the study protocol.

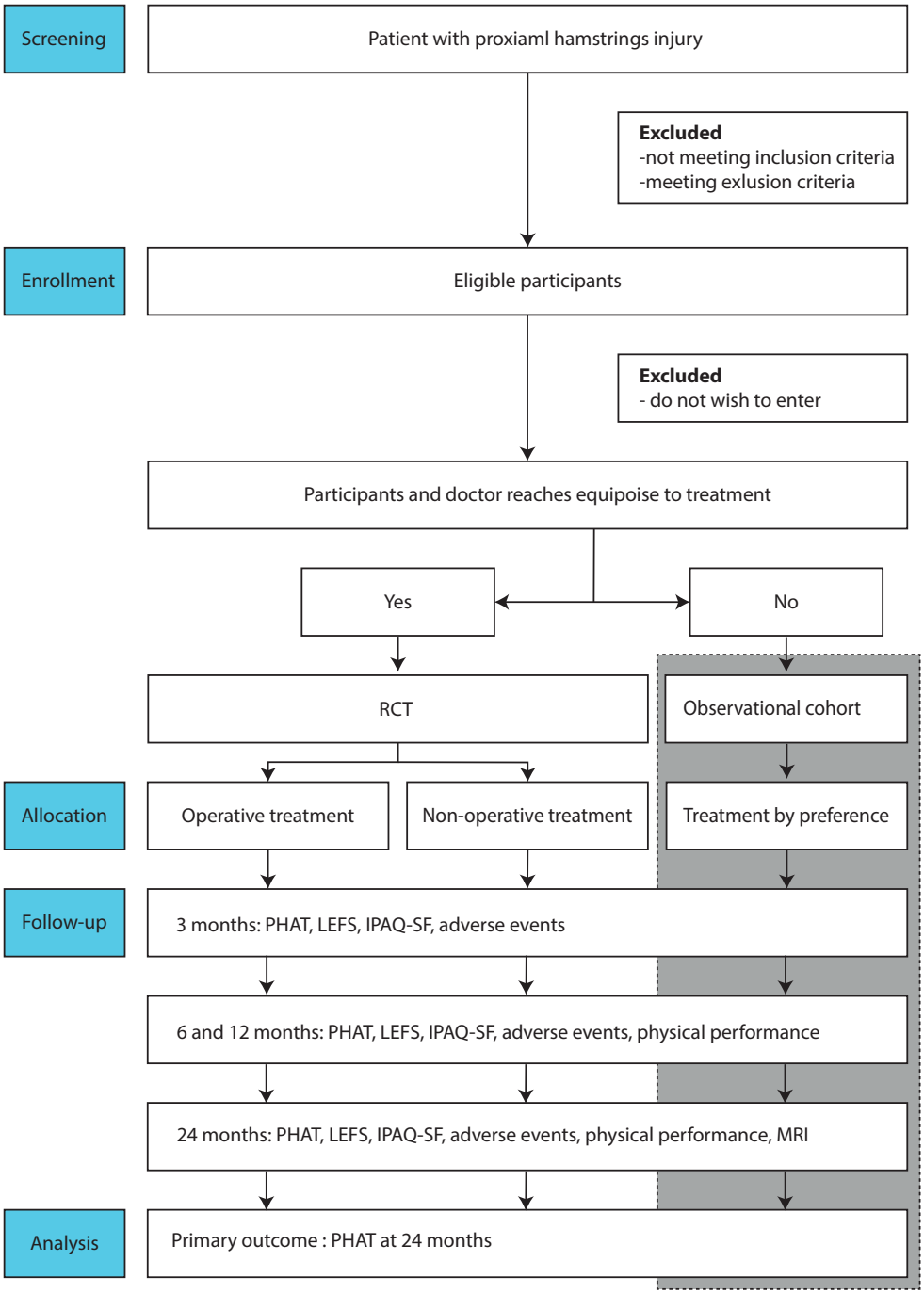

Figure 1. *Study design. Patients aged 30-70 years old with proximal hamstring avulsions who meet inclusion criteria and none of the exclusions criteria are invited for participation. IPAQ-SF, International Physical Activity Questionnaire Short Form; LEFS, Lower Extremity Functional Scale; PHAT, Perth hamstring assessment tool; RCT, Randomized Controlled Trial. Patients are followed for 24 months with study visits at 3, 6, 12, and 24 months.*

4.2 Endpoints

4.2.1 Primary

The primary endpoint (measured at baseline, 3, 6, 12 and 24 months) is self-reported Perth hamstring assessment tool (PHAT). The primary analysis will be conducted using the measurement at 24 months.

4.2.2 Secondary

The secondary endpoints can be grouped into
a) adverse events and complications (measured at 3, 6, 12 and 24 months)
b) additional patient reported functional outcome scores (measured at baseline (if applicable), 3, 6, 12 and 24 month)
c) physical performance-based tests (measured at 6, 12 and 24 months) including
    c1) strength tests and
    c2) functional tests
    c3) range of motion
e) imaging outcomes (measured at 24 month).

4.2.2.1 Adverse events and complications (both groups)

- Surgical site infections (surgery group only)
- Neurological sequel
- Thromboembolic disease
- Re-rupture in surgically treated patients
- Other complications and reoperations

4.2.2.2 Additional patient reported functional outcomes

- The lower extremity functional scale (LEFS; see protocol appendix LEFS).
- Self-reported pain score at rest, during sitting and during walking in the groups. This is a subset of the PHAT score (see protocol appendix PHAT).
- Patients reporting that they have returned to preinjury sporting activities (see protocol appendix return to sports).
- Visual Analogue Scale (VAS) satisfaction of treatment (see protocol appendix satisfaction).
- VAS self-reported recovery (see protocol appendix satisfaction).
- Activity level measured by IPAQ-SF preinjury level to follow-up (see protocol appendix IPAQ-SF).
- Time to return to work.

4.2.2.3 Physical performance-based tests of both injured and uninjured leg.

Hamstring muscle strength

Maximum isometric force in supine position (see protocol appendices physio protocol and instruction). Study sites equipped with a computer-based isokinetic dynamometer will test maximum isokinetic force during knee flexion and hip extension (see protocol appendices physioprotocol and physioinstruction).

Functional tests

- Timed step test (see protocol appendices physio protocol and physio instruction) of the injured and uninjured side.
- Single leg hop tests (see protocol appendices physio protocol and physio instruction) of the injured and uninjured side.

Range of motion

Hip flexion and knee extension (see protocol appendices physio protocol and physio instruction)

4.2.2.4 Radiological

- Hamstrings muscle volume and fatty infiltration in the injured and uninjured side measured by MRI at 24 months (see protocol appendix MRI protocol 24 months PHACT).
- Attached tendons at the ischial tuberosity measured by MRI at 24 months.

These endpoints are measured at 24 months.

4.3 Randomization and blinding

The REDCap (REDCap Software) randomization tool will be used to perform randomization of patients, which will be conducted using a random block size (2-6) stratified by study site.

To minimize ascertainment bias this trial is single-blinded, where the physiotherapist conducting strength and functional tests at 6, 12 and 24 months will be blinded to the intervention, by informing the patients not to tell what group they belong to and asking them to wear clothes concealing potential surgery scar. Statisticians analyzing the data will also be blinded to treatment arms. Consequently, analyses of adverse events will be conducted last (since these are likely to reveal group allocation). Similarly, results will be presented and discussed among investigators prior to analyzing and presenting adverse events (to preserve blinding as long as possible in the construction of the first manuscript from the trial).

4.4 Data management

All study data will be collected and managed in a digital case report form using REDCap electronic data capture tools hosted at Karolinska Institutet, Sweden. REDCap is a secure, web-based application supporting data capture for research studies.

Data will be kept securely in order to protect confidentiality before, during, and after the trial. A codebook matching the personal identification number and the trial identification number is kept at each study site and the trial identification number is noted in the patient's electronic medical chart. The study nurses and investigators can log on and enter data directly into the database. Patients will complete surveys at each visit. Any paper forms used are stored for cross-checking at each study site.

4.5 Trial reporting

The trial will be reported according to the principles of the CONSORT statements in scientific publications.

5 Analysis of the trial

5.1 Analysis populations

The *intention to treat (ITT) population* will consist of all participants randomized and consenting to be part of the study, irrespective of treatment received. Participants in the ITT population will be analyzed according to the treatment they are randomized to, regardless of whether they actually received this treatment or not.

We will also conduct *per-protocol (PP)* and *as-treated (AS)* analyses. Cases will be considered treatment crossovers if the randomly assigned treatment is changed. Non-operative treated patients who are treated operatively due to late complaints (>3 months after inclusion) will in PP and AS analyses be handled in two ways: (i) *not* considered crossovers, in which case we regard these sets of patients as a failed initial conservative treatment strategy, followed by surgery; (ii) considered as cross-overs (analogously to patients who switch groups early, i.e. within 3 months of randomization). The first definition will be employed in the main PP and AS analyzes, whereas the latter will be used in sensitivity analyses (see Section 5.8.4).

The *per protocol (PP) population* will be defined by the participants who were randomized to a specific treatment and received it according to the protocol. Analyses based on the PP population are based on post randomization events (treatment compliance vs. not), which may lead to biased results. We will therefore perform analyses on the PP population using inverse probability weighting to adjust for the effect of these post randomization events. Specifically, let A be an indicator variable that takes value 1 if the patient was randomized to the surgical arm and 0 If s/he was randomized to the conservative treatment arm. The conditional probability of adhering to the protocol given study arm and other covariates is $Pr(P=1|Z,A)$, where Z is patient covariates (age, sex, site, and degree of tendon retraction). Inverse probability weights of protocol adherence will be defined as $W=1/Pr(P=1|Z,A)$. The weights W will then be used to estimate the contrasts between the operative and non-operative arms using marginal models with a robust (sandwich) standard error estimator according to models specified in Sections 5.2 and 5.3.

The *as treated (AS) population* will be defined according to the treatment the patients actually received. Analogously to the PP population, the definition of the AS population is based on post randomization events, which may lead to biased results. We will therefore conduct analyses on the AS population by first identifying subgroups that are likely to require surgery and therefore should not be considered as good candidates for conservative treatment, and then contrast the outcomes of patients treated successfully in the conservative treatment arm with patients randomized to the surgery arm who similarly would not have crossed over had they been randomized to the conservative treatment arm. We will within the trial observe which patients in the conservative management arm who cross over. However, patients cannot cross over from the surgical to the conservative arm, after the surgery is performed. We will use matching based on inverse probability weighting (as defined above for the PP population) to identify a comparison group from the surgical arm for compliers in the conservative management arm.

The number and proportion of patients who did not receive the treatment they were allocated to will be reported. If any other trial treatment options were known to be received, instead of or in addition to, the allocated treatment during the follow-up period after the first randomization, these will be collected and reported.

5.2 Primary endpoint

PHAT scores (measured at 3, 6, 12, and 24 months) will be assessed and contrasted between the two treatment groups using a linear regression model that adjusts for an indicator of randomized treatment group assignment and for factors used to stratify randomization (site), age, sex, and degree of tendon retraction. The primary analysis will contrast the randomized groups at 24 months in the ITT population (analyses using other follow-up time points are specified in Section 5.8).

Based on the existing literature the standard deviation of PHAT measurements is ~16 – 21. The non-inferiority margin was set to half of the standard deviation (=10). The non-inferiority margin was agreed upon at a consensus group meeting that included orthopedic surgeons and statisticians. The one-sided alpha level will be 0.05. If conservative treatment is non-inferior on the one-sided 0.05 alpha level, we will repeat the test using a one-sided alpha of 0.025.

5.3 Secondary endpoints

The main analyses of the secondary endpoints will contrast the randomized groups at 24 months (analyses using additional time points are specified in Section 5.8).

5.3.1 Adverse events and complications

The following adverse events will be collected:

- Surgical site infections
- Neurological sequel in both groups
- Thromboembolic disease
- Re-rupture in surgical treated patients

Adverse events (AEs) will be summarized in tables by arm. If justified based on the number of AEs in each arm, we may also analyze the safety endpoints in two different ways: (1) Any AE using a logistic regression model with arm allocation as the main contrast, adjusted for site, age, sex, and degree of tendon retraction; (2) Total number of AE using Poisson regression with arm allocation as the main contrast, adjusted for site, age, sex, and degree of tendon retraction.

5.3.2 Additional patient reported functional outcomes

1. The difference between the groups in the LEFS.
2. The difference between self-reported pain score at rest, during sitting and during walking in the groups (subset of the PHAT score).
3. Difference between the groups in the percentage of patients that report that they have returned to preinjury sporting activities.
4. Difference in VAS Satisfaction of treatment.
5. Difference in VAS self-reported recovery
6. Difference in in activity level measured by IPAQ short

Endpoints 1, 2, 4, and 5 in the list above will be analyzed in linear regression models analogously to the analysis of the primary endpoint. Endpoint 3 in the list above will be analyzed in a logistic regression model (adjusted for site, age, sex, and degree of tendon retraction) at each follow-up time point.

5.3.3 Hamstring muscle strength

The difference between the two treatment groups in the ratio between the injured and uninjured side of the maximum isometric strength in supine position and isokinetic force during knee flexion and hip extension (see protocol appendix physioprotocol and appendix physioinstruction). We will test for differences in the ratio between the groups using a Wilcoxon test, for ease of interpretation. We may also analyse the difference between the groups in a linear regression model, where the dependent variable will be muscle strength in the injured leg, and the independent variable will be muscle strength in the uninjured leg at baseline and the relevant follow-up timepoint, the muscle strength in the injured leg at baseline, site, age, sex, and degree of tendon retraction.

5.3.4 Physical performance-based tests

- The difference between the two treatment groups in the mean ratio between the injured and uninjured side of the timed step test.
- The difference between the two treatment groups in the mean ratio between the injured and uninjured side of the single leg hop tests.

These endpoints will be analysed similarly to the analysis for hamstring muscle strength (see Section 5.3.3).

5.3.5 Range of motion

The difference between the two treatment groups in the ratio between the range of motion in knee extension and hip flexion. This endpoint will be analysed similarly to the analysis for hamstring muscle strength (see Section 5.3.3).

5.3.5 Radiological outcomes

- Difference between groups in the ratio of hamstrings muscle volume and fatty infiltration between the injured and uninjured side measured by MRI at 24 months.
- Ratio of attached tendons at the ischial tuberosity measured by MRI at 24 months.

These endpoints will be analysed similarly to the analysis for hamstring muscle strength (see Section 5.3.3).

5.5 Descriptive analyses

5.5.1 Trial flowchart

The flow of participants through the trial will be summarized using a CONSORT diagram. The flow diagrams will describe the numbers of participants randomly allocated, who received allocation, withdrew consent, and included in the ITT, PP, and AS analysis populations. Also the observational cohort may be depicted in the flowchart.

5.5.2 Baseline characteristics

Baseline characteristics will be described descriptively and will include:

- Age
- Sex

- Injured leg
- BMI
- Occupational level
- Activity at injury
- PHAT
- LEFS
- IPAQ-SF
- Degree of tendon retraction
- Number of tendons detached (conjoint, semimembranosus or both)
- Time to treatment, (days)

Summary statistics will be used for each variable: median and interquartile ranges for continuous variables, and number and percentages for categorical variables.

5.5.3 Trajectories – display of results over time

We may produce graphs to show the trajectory of individual patients and the randomized arms (as well as the observational cohort) with respect to the follow-up of endpoints (primary as well as secondary) over time.

5.6 Subgroups

Results may be stratified by the following subgroups:

- Site
- Age
- Sex
- Degree of tendon retraction
- Number of tendons detached (conjoint, semimembranosus or both)

5.7 Missing data

All reasonable efforts will be taken to ensure that the level of missing data and loss to follow-up will be minimal.

Missing data can occur in different ways in the study: (1) individual questions in the instruments (e.g. the PHAT questionnaire) can be left unanswered; (2) data on specific endpoints may be entirely missing at different follow-up timepoints (due to a e.g. a functional test not being performed or data on a specific instrument not collected); and (3) patients can miss specific follow-up visits or drop out of the study altogether (resulting in no information on endpoint data for the corresponding follow-up timepoint or missing data on all subsequent follow-up timepoints).

Missing data on individual instrument (e.g. PHAT) questions will be imputed using predictive mean matching (PMM), with donor pool (k)=5.

Missed follow-up visit at specific follow-up timepoints will be handled using a multiple imputation by chained equations (MICE). The multiple imputation protocol will be based on models for predicting outcomes at specific follow-up timepoints based on outcomes data recorded at other time points together with treatment group, patient age, sex and degree of tendon retraction. For example, a missing PHAT score at 24 months will be imputed based on a regression model fit to

data from timepoints 3, 6, and 12 together with treatment group, patient age, sex and degree of tendon retraction.

We will generate 1,000 datasets with imputed outcome data, which will be analyzed separately. We will then use Rubin's rules to pool the estimated absolute differences in PHAT score and standard errors.

The global COVID pandemic has occurred during the conduct of this trial. COVID and restrictions have had a major impact on the healthcare situation and has resulted in rescheduled and cancelled follow-up visits in PHACT. Therefore, visits will be analyzed as follows:

- Visits occurring 2.5 to 4.5 months after randomization (or baseline for the observational cohort) will be considered as the 3 month visit.
- Visits occurring 4.5 to 9 months after randomization (or baseline for the observational cohort) will be considered as the 6 month visit.
- Visits occurring 9 to 18 months after randomization (or baseline for the observational cohort) will be considered as the 12 month visit.
- Visits occurring >18 months after randomization (or baseline for the observational cohort) will be considered as the 24 month visit. If there are several such visit, the one closest to the 24 month mark will be used.

Since the date for PROM reporting, MRI, and physiotherapy visits may differ, we will use the data if the primary endpoint (PHAT) reporting for the grouping.

5.8 Additional analyses

5.8.1 Heterogeneous treatment effects

We will test for heterogeneity of treatment effects by testing for significant interactions (following the best practices described in Wang et al. N Engl J Med 2007; 357:2189-2194) in the following subgroups: tendon retraction >2 versus ≤2 cm and age >50 versus ≤50 years and IPAQ score above or below the median IPAQ at baseline (median computed across the two randomized arms).

5.8.2 Analyses including different follow-up time points

The analyses specified for the primary and secondary endpoints in Sections 5.2 and 5.3 at 24 months of follow-up may be replicated at 3, 6 and 12 months of follow-up (i.e. analogous regression models will be used, but using endpoint data from 3, 6 and 12 months of follow-up instead of 24 months).

We may also jointly analyze all time points in a linear mixed model (to adjust for within-patient correlations). Patients will be treated as a random effects, and time points, randomization arm, age at baseline, sex and degree of tendon retraction will be included as fixed effects. Mixed effect models corresponding to analogous fixed effect model for individual time points will be used (i.e. a logistic mixed effect model where a logistic model was used for an individual time point, etc.). The mixed-effects model handles data missing at random, however drop-out processes can be missing not at random (MNAR). Therefore, we may use MICE also in the context of linear mixed model analyses. The MICE procedure will then be done analogously to what is described in Section 5.7.

5.8.4 Sensitivity analysis with respect to definition of cross-overs

As specified in Section 5.1, non-operative treated patients who are treated operatively due to late complaints (>3 months after inclusion) will in PP and AS analyses be handled in two ways: (i) *not* considered crossovers, in which case we regard these sets of patients as a failed initial conservative treatment strategy, followed by surgery; (ii) considered as cross-overs (analogously to patients who switch groups early, i.e. within 3 months of randomization). We may perform sensitivity analyses using both these definitions of a cross-over from conservative to surgical treatment.

5.8.4 Analyses including the observational cohort

The randomized and observational cohorts will be analyzed together propensity scores. The propensity score will be based on a logistic regression model including age, sex, study site, baseline IPAQ and the degree of tendon retraction as covariates. The propensity scores will be used to perform inverse probability of treatment weighting (IPTW) with stabilized weights. The propensity score is defined as $e = P(A = 1|Z)$, where A is treatment allocation (A=1 for surgical treatment and A=0 for conservative treatment) and Z are the covariates. The stabilized weights are then defined as $W=A*Pr(A=1)/e+(1-A)*Pr(A=0)/(1-e)$.

We may also perform analyses using the observational cohort that are analogous to the ones we will perform to contrast the two randomized arms, but instead contrasting the observational cohort with the randomized cohort or one of the randomized arms.

5.8.5 Sensitivity analysis of the primary endpoint

The PHAT score is constrained between 0 and 100. For the primary analysis, we use linear regression, which is not constrained to values between 0 and 100. As a sensitivity analysis, we may use beta regression instead of linear regression to analyze the primary endpoint (using the same covariates), since beta regression is bounded to an interval between 0 and 1. Using the fitted beta regression model, we can compute the difference in PHAT score using regression standardization to marginalize across the covariate distributions.

5.8.6 Sensitivity analysis of adjustment for site

Site is a stratification variable in the randomization. As such, it is advisable to adjust for site in the analysis of the primary and secondary endpoints. However, it is not unlikely that some sites will recruit a small number of patients. Adjusting for many small sites raises analytical problems for which there is no best solution. Analyses either ignoring site or adjusting for a large number of small sites might lead to confidence intervals that may be either too large or too small. Further, pooling small sites has no scientific justification. We will approach this problem as described below:

- If all sites have ≥10 patients included in the randomized cohort, we will adjust for site in the analyses. We may then conduct sensitivity analyses where site is not adjusted for to assess its impact on the results.
- If at least one site has <10 patients included in the randomized cohort, we will not adjust for site in the analyses. We may then conduct sensitivity analyses where site is adjusted for to assess its impact on the results.

6 Sample size

6.1 Original sample size calculations

In PHACT, the PHAT-score is used as the primary outcome measure. It has been previously shown that minimal detectable difference in PHAT is 16 points and SD about 15 to 20. The sample-size calculations assumed a noninferiority margin of 16, a one-sided alpha of 2.5, 80% statistical power, and a standard deviation of the PHAT score of 20, which yielded a required sample size per group of 25 patients. Assuming 20% drop-out rate, we aimed at a group size of 30 (total 60). As cross-over was expected, we also decided to continue recruitment until 30 patients in each arm had entered and initiated the allocated treatment.

6.2 Updated sample size calculations (181204)

The study progress and accrual are continuously monitored. Specifically, we have pre-specified an evaluation of the study after 60 included participants with a possible increase in the size of the study as a result of the evaluation if it is motivated by a greater dropout than expected or new vital information obtained from discussions with patients and clinical staff. The study has in December 2018 been ongoing for 14 months and about 50 patients have been included in the study. It is clear that our original estimate of the incidence of the injury was too low. We also conclude that the logistics and operational aspects of the study are well-functioning. Thus, we have the possibility to consider an increase in the size of the study. The study has a unique opportunity to answer not only the primary question but also to analyze the effect of age and the severity of the injury on the effect of the two treatments.

To do this, we have discussed the ethical aspects of such a change:

Pros:
• A larger number of patients in the study increases the power of the study; i.e. reduces the risk of type 2 errors.
• A larger number of patients provides a better basis for subgroup analyzes. This is something that in conversations with patients has emerged as a crucial aspect, as a specific patient is interested in the probable outcome for him or her (i.e. patients who have the same characteristics) rather than the overall result at the population level.

Cons:
If the study in its current form can deliver a clear result that affects the treatment of future patients, the dissemination of result will be delayed by the study becoming larger and a number of patients will be treated without the new knowledge being reported. We believe that this risk is limited and justified. Based on clinical experience and existing scientific literature, there are no indications that the treatment result differs clearly between the groups and a larger study size provides a greater opportunity for results that are conclusive and lead to a change in clinical practice.

In the light of the description and motivation above, we decided to make the following change to the trial: In the literature, surgical treatment is strongly advocated. We wish to analyze the study in such a way that we want to be sure that the non-surgical treatment is not worse than the surgical treatment with a certain margin (non-inferiority trial with respect to the PHAT score).

To achieve 85% power, with a one-sided alpha of 0.05, for demonstrating non-inferiority using a non-inferiority margin of 10, 50 patients in each arm are required (assuming a standard deviation of the PHAT score of 20; which in the scientific literature has been reported to be in the range of 16 to 21). Accounting for drop-out, we will include 60 participants in each trial arm in the randomized part of the trial. Table 1 gives estimated power for 60 patients in each trial arm (120 patients in total) under different assumptions and alpha levels.

N	Dropout (%)	SD	Alpha (one-sided)	Power (%)
120	0	20	0.05	85
120	0	20	0.025	77
120	10	20	0.05	82
120	10	20	0.025	74
120	20	20	0.05	79
120	20	20	0.025	69
120	0	16	0.05	95
120	0	16	0.025	92
120	10	16	0.05	93
120	10	16	0.025	90
120	20	16	0.05	92
120	20	16	0.025	86

7 Post-hoc analyses

7.1 Estimates of relative risks for binary endpoints

Logistic regression was used for binary secondary endpoints (specifically, adverse events and returning to sports). Since odds ratios are estimated from logistic regression models and since odds ratios may overestimate relative risks, we performed a post-hoc analysis to estimate relative risks for the adverse events and returning to sports secondary endpoints. This was performed according to the following: Marginal relative risks to compare the two treatment arms were calculated starting from a logistic regression including treatment as a covariate and the binary endpoint as the dependent variable, adjusted for age, sex and degree of tendon retraction. Marginal estimates presented are population-averaged adjusted risk ratios and the values are obtained by taking the average of unit-level estimates. Specifically, the marginal relative risks were then calculated by the ratio of the mean predicted risk if all patients were treated nonoperatively to the mean predicted risk if all patients were treated operatively.

7.2 Limb symmetry index at 24 months

In a post-hoc analysis, we computed the confidence interval for the limb symmetry index ([value of injured side/value of uninjured side]*100) for the hamstring muscle volume and fat fraction (measured using MRI) at 24 months *within* the operative and nonoperative groups for both the RCT and observational cohort. Since the LSI is a ratio of two correlated random variables, we used the nonparametric bootstrap (with 10,000 samples) to compute 95% confidence intervals.

7.3 Changes to the SAP as a result of the review process

As part of the review process, two main changes were made to the analysis of the trial. Specifically:

1. Instead of using the ITT population for the primary analysis, the PP population was used.
2. Instead of performing analyses where the RCT cohort and the observational cohort were analyzed together, separate analyses were conducted and reported for two cohorts.