DIAGNOSTIC ACCURACY OF THE THESSALY TEST, THE STANDARDISED CLINICAL HISTORY, AND OTHER CLINICAL EXAMINATION TESTS FOR MENISCAL TEARS

STATISTICAL ANALYSIS PLAN

1. INTRODUCTION

1.1. STUDY BACKGROUND

The menisci are two semilunar, fibrocartilaginous disks located between the medial and lateral articular surfaces of the femur and tibia in each knee. The menisci play an important role in the function of the knee providing load bearing, stress distribution and shock absorption across the knee. Tears in the menisci are a common knee injury that can cause pain in the joint. In younger active patients tears are often a result of sports injuries. In older people degenerative meniscal tears are more common. Reliable non-invasive diagnosis of meniscal tears is difficult. There are a number of physical examination tests that diagnose tears but all suffer a lack of specificity (the correct identification of those that do not have a meniscal tear) and sensitivity (the correct identification of those that do have a meniscal tear).

MRI is often referred to as the gold standard for non-invasive diagnosis of meniscal tears. However, incidental meniscal findings on MRI of the knee are common in the general population, increase with age and may not be associated with pain. Meniscal damage is also a frequent finding on MRI of the osteoarthritic knee limiting the value of this diagnostic tool for meniscal tears in this section of the population.

The Thessaly test is a clinical examination used to detect miniscal tears in the knee. Established alternative tests to the Thessaly test include the McMurray test, Apley's test and joint line tenderness test. Previous reports have come to the conclusion that a combination of tests is required to produce accurate diagnoses.

The accuracy of specialist knee clinicians in performing physical examinations of the knee may differ to primary care staff who will inevitably see fewer patients with knee pathology and have less training in performing tests.

1.2. STUDY OBJECTIVES

The objectives of the study are:

- to determine the diagnostic accuracy of the Thessaly test by GPs for meniscal tear in the knee and whether this test can obviate the need for further investigation by arthroscopy or magnetic resonance imaging (MRI);
- to determine how the Thessaly test compares to clinical history and other commonly used physical examinations (McMurray test, Apley's test, joint line tenderness test) in diagnosing meniscal tears by GPs;
- to determine if the presence of arthritis or other knee pathologies influences the accuracy of the Thessaly test;
- to determine if the use of combinations of physical tests (such as the Thessaly test, McMurray test, Apley's test and or joint line tenderness test) by GPs provides better specificity and sensitivity than a single test alone in the diagnosis of meniscal tear;
- to determine the ability of non-specialist General Practitioners (GPs) to use the Thessaly test in comparison to specialist knee clinicians.

1.3. STUDY DESIGN

This is a single centre (Glasgow Royal Infirmary) observational diagnostic study. 300 patients will be attending knee clinics at Glasgow Royal Infirmary, and have suspected knee pathology. 5-10% of this group will be enrolled via a single

general practice, and will be all patients presenting at the GP with knee symptoms, and will be sub-analysed to check comparability of the wider group with the primary care population. 50 patients will be attending orthopaedic hand clinics, and have no suspected knee pathology, acting as controls.
All participants (who will attend weekly knee clinics at the Glasgow Royal Infirmary) will be assessed using the Thessaly test, McMurray test, Apley test and joint line tenderness test, independently by orthopaedics specialist clinicians and GPs. The order of these 4 tests will be randomly permuted. Likewise if feasible the order of specialist clinician and GP. All participants with will undergo MRI scan and knee x-ray (to identify the subgroup of patients with arthritis in the knee; control subjects will not have knee x-ray). All participants will have a medical history taken (with half randomly assigned to take the medical history before the tests, half after the tests). Arthroscopy will be performed only on patients who would normally receive this as part of their standard care.
There will be 3 specialist orthopaedic clinicians and 10 general practitioners. Each patient will be assessed by one specialist orthopaedic clinician and one GP. It is expected that each specialist orthopaedic clinician and each GP will assess roughly equal numbers of patients.
The GPs and specialist orthopaedic clinicians will be unaware of each others test results and also the referent gold standard MRI test and the X-ray test to establish arthritis in the knee.

1.4. Sample Size and Power

The following sample size justification is given in the study proposal:

> Assuming the sensitivity of the Thessaly test is around 75%, the study would need around 300 subjects to estimate the sensitivity to within +/- 5%. A similar calculation for the width of the confidence interval for a binomial proportion is appropriate for the specificity – for example, if the specificity was around 90%, the required sample size to estimate the specificity to within +/- 8% would be n=50 participants. The power for the pairwise comparison of tests, or combinations of tests, will depend on the degree of disagreement between the tests – for example, with around 220 pairs of measurements the study would have 90% power to detect a difference in proportions of 0.10 when the proportion of discordant pairs is expected to be 0.15 (using McNemar's test).

1.5. Study Population

1.5.1. Inclusion Criteria

Knee pain group (N=300):

- Patients referred to the knee clinic at Glasgow Royal Infirmary.

Control group (N=50):

- Patients attending the hand clinic at Glasgow Royal Infirmary.

1.5.2. Exclusion Criteria

Knee pain group (N=300):

- Age under 18;
- Unable to give informed consent;
- Previous knee replacement.

Control group (N=50):

- Age under 18;
- Unable to give informed consent;
- Previous knee surgery;
- History of knee pain (last 6 months);
- Osteoarthritis;

- Rheumatoid arthritis.

1.6. Statistical Analysis Plan (SAP)

1.6.1. SAP Objectives

The objective of this SAP is to describe the statistical analyses to be carried out for the study titled "Diagnostic accuracy of the Thessaly Test, the standardised clinical history, and other clinical examination tests for meniscal tears".

1.6.2. Current Protocol

At the time of writing, no formal study protocol has been written. This document is based on the study proposal, submitted as a full proposal to the NIHR HTA Commissioning Board in September 2010, and approved for funding in June 2011. Future development of the protocol will inform subsequent versions of this SAP, which will be updated as necessary.

1.6.3. General Principles

Results will be presented for the study population as a whole and separately for the knee pain and control groups. Within the knee pain group, results will also be presented separately for the subgroup of patients referred from a single general practice in comparison to all other patients, for those with and without arthritis, and in other subgroups according to knee pathology (Anterior Cruciate Ligament (ACL) rupture, or other previous injury or treatment of the knee) and patient characteristics (including the subgroup of predominantly younger patients with sports injury and the subgroup of those with degenerative changes due to age).

Diagnostic performance measures will be calculated for each individual test (Thessaly test, McMurray test, Apley test, joint line tenderness test and clinical history), using evidence of meniscal tear on MRI as the referent (gold standard) test. It is recognised that this MRI gold standard is itself imperfect. However, it is the established best diagnostic tool available on which intervention and treatment decisions are made, and no feasible alternative exists or is available. The primary interest will be the performance of these tests when used by GPs. Results will be reported for the tests performed by Orthopaedic Clinicians, and compared to the performance achieved by GPs.

Combinations of physical tests will be considered, to determine the optimal combination for the diagnosis of meniscal tear. Logistic regression methods will be used to determine whether the addition of patient characteristics to the results of physical tests provides greater discriminatory ability.

1.6.4. Software

Statistical analyses will be carried out using SAS for Windows v9.2, R for Windows v 2.12.1 or SPlus for Windows v8.1, or higher versions of these programs.

2. Analysis

2.1. Study Populations

The numbers of people screened and the numbers and percentages recruited in each study population will be presented, as will the numbers providing data for each diagnostic test. Numbers of participants not completing the study according to the protocol, with reasons for non-completion, will be presented.

2.2. Baseline Characteristics

Summary tables will be presented, describing the baseline characteristics of each study population. Appropriate statistical tests will be used to compare the different populations. Similar summaries and tests will be used to describe population subgroups of particular interest.

2.3. Diagnostic Test Results

2.3.1. Descriptive Statistics

The numbers and percentages of individuals classified as having meniscal tears according to each test will be presented for each study population and in subgroups of particular interest. Results of physical tests performed by GPs and Orthopaedic Clinicians will be presented separately and compared with exact McNemar tests.

2.3.2. Single Tests

The diagnostic properties of the tests will be summarised using standard techniques for diagnostic studies as described by Pepe, 2003.

The sensitivity (Sens) and specificity (Spec) of each physical test will be presented, along with the positive and negative likelihood ratios (LR+ and LR-), and the diagnostic odds ratio (DOR) with exact 95% confidence intervals, treating the MRI result as the true diagnosis. This is for the group with knee symptoms. For the controls from the hand clinic, we do not expect any positive meniscal tear diagnosis on MRI – for this group the objective is to compare specificities.

Likewise, we will also calculate the positive predictive value (PPV) and the negative predictive value (NPV) of the 4 simple tests, along with the appropriate exact 95% confidence intervals, to summarise what a positive and negative test tells us in those that have knee symptoms, assuming this represents the population of patients in primary care with suspected meniscal tear. The appropriateness of the assumption will be assessed by comparison with the subgroup of consecutive patients with knee symptoms referred from a single general practice, and consideration given to reweighting the estimates via an appropriate statistical model to account for any systematic differences between the two populations, if necessary. The characteristics compared will include age, gender, socioeconomic status, and various medical history items. These will be compared using t-tests and chi-squared tests as appropriate.

We will plot the Sensitivity vs. 1 – Specificity for the 4 simple tests in the group with knee symptoms to visualise their relative performance.

The performance of tests performed by GPs and Orthopaedic Clinicians will be presented separately. The main interest is in the performance of the GPs. The performance of these tests at the hands of the specialist orthopaedic clinicians is expected to indicate an upper bound on their potential performance. The GP and specialist orthopaedic clinicians performance will be compared with exact McNemar tests.

Physicians' views on the use of the different physical tests will be summarised and compared between tests.

2.3.3. Combined Tests

The diagnostic performance of alternative combinations of physical tests will be estimated. We will use various methodological approaches as discussed in Knotterus, 2009:

- Logistic regression, with MRI classification (meniscal tear, Yes/No) as the outcome, will be used to build a series of models on the GP's performance to assess the diagnostic properties, as follows:
    - Core model: including 'design' information (indicator variables for the randomised order of the tests, randomised order of taking the medical history), and GP as a random effect.
    - Model Level 1: The Core model with an individual test in isolation (4 models)
    - Model Level 2: The Core model with participant baseline covariates (age, sex, previous history, socioeconomic status, and so on) (1 model)

- Model Level 3: Re-do model Level 1 with Model level 2 covariates (4 models)
- Model Level 4: Explore GP characteristics as influences e.g. age (or time since qualified), gender, specialities, GP status (e.g. partner), GP surgery characteristics (e.g. number of partners), GP practice size, and so on.
- Model Level 5: Stepwise selection model to establish parsimonius model combining GP and patient level predictors to provide Updated Core Model
- Model Level 6: Investigation of combinations of pairs of 2, triplets of 3 and all 4 tests combined in the presence of the Updated Core Model.

All the models will be assessed by their concordance index (c-statistics) measuring the area under the curve. When considering whether an increment in the c-statistic moving from one model to the next is worthwhile, due allowance will be made for the increased complexity of the model.

We will consider adding in the patient-defined subgroups (such as arthritis (yes/no), ACL rupture (yes/no), sports injury (yes/no), degenerative disease (yes/no) as subgroups of particular interest in the development of these models, and formally test for interactions as appropriate.

We will consider re-running this modelling hierarchy for the specialist orthopaedic clinicians data.

.

- Classification And Regression Trees (CART, as implemented in R) will also be used to determine an optimal combination of tests provides better prediction. The advantage of this approach is that it allows complex interactions between the four tests which the logistic regression approach isn't naturally suited for. The disadvantage is that CART is purely data driven, and hence often produces solutions which do not transfer to the next dataset. We will look at 'averaging' trees across split samples using resampling techniques to try to overcome this and produce stable, robust trees.

  As with the logistic regressions, we may will investigate specific subgroups of interest and possibly re-run on the specialist orthopaedic clinicians dataset.

2.3.4. REPORTING

The study will be reported to the standards established in the STARD initiative (Bossuyt et al, 2003)

2.3.5. MISSING DATA

We do not anticipate any missing data arising from any issues regarding the co-operation or availability of the specialists clinicians or the GPs – the sessions at which the measurements will be taken will be arranged in advance to suit these health professionals. It may happen that not all 300 of those with knee symptoms or all the 50 non-knee symptom controls are not available for all measurements – we will endeavour to make sure we reach these targets. In terms of baseline covariate measurements on both participants and clinicians, these are all simple information and again we do not anticipate not having full information on everyone. As such it is not anticipated that missing data will be an important issue in this study, so we will simply describe what if anything is missing and we have no special plans for dealing with missing data in the analyses.

3. References

Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. OUP (2003).

The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. Editors Knottnerus JA, Buntrix ID. Chapter 8 – Multivariate analyses in diagnostic accuracy studies: what are the possibilities? BMJ Books, Wiley-Blackwell (2009)

Bossuyt PM, Reitsma JB, Bruns DE Gatsonis PP, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, De Vet HC; Standards for Reporting of Diagnostic Accuracy. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. BMJ (2003); 4; 326; 7379; 41-4