

# **The Diagnostic Accuracy of the Clinical Examination of the Hip**

Steven Gordon White

A thesis submitted to AUT University in fulfilment of the requirements for the degree of  
Doctor of Philosophy (PhD)

2016

School of Clinical Sciences

Supervisors

Professor Peter McNair

Dr Mark Laslett

## Table of Contents

Table of Contents.....	i
List of Figures.....	v
List of Tables.....	vi
Attestation of Authorship.....	viii
Research Outputs Resulting from this Thesis.....	ix
Acknowledgements.....	x
Ethics.....	xii
Abstract.....	xiii
Chapter 1 Introduction.....	16
1.1 The problem.....	16
1.2 Research aims.....	18
1.3 Structure of the thesis.....	20
1.4 Significance of the research.....	21
Chapter 2 General Literature Review.....	22
2.1 Introduction.....	22
2.2 General search strategy.....	23
2.3 Intra-articular hip joint pain.....	23
2.3.1 Pain.....	24
2.3.2 Anatomy.....	26
2.3.3 Innervation of intra-articular structures of the hip.....	27
2.3.4 Review Summary.....	33
2.4 The diagnostic process.....	34
2.5 Performance characteristics of tests.....	37
2.5.1 Sensitivity and specificity.....	38
2.5.2 Predictive values.....	39
2.5.3 Likelihood ratios.....	40
2.5.4 Overall accuracy.....	43
2.5.5 Review Summary.....	46
2.6 Chapter summary.....	47
Chapter 3 Pain Provocation and Pain Intensity During the Application of Physical Tests.....	48
3.1 Introduction and background.....	48
3.2 Literature review.....	49
3.2.1 Intra-examiner reliability of reports of pain reproduction and pain intensity.....	50
3.2.2 Inter-examiner reliability of reports of pain reproduction.....	51
3.2.3 Prevalence of painful responses during the application of pain provocation tests.....	52
3.2.4 Methodological considerations.....	54
3.2.5 Review Summary.....	57
3.3 Methods and procedures of the current study.....	58
3.3.1 Study design.....	58
3.3.2 Sample size calculation.....	58
3.3.3 Participants.....	58
3.3.4 Examiner.....	59
3.3.5 Procedures.....	59
3.3.6 Blinding.....	61
3.3.7 Included tests.....	61

3.3.8	Definition of a positive test .....	61
3.3.9	Analysis .....	61
3.4	Results .....	63
3.4.1	Reliability of categorical yes/no response to pain reproduction .....	67
3.4.2	Reliability of reports of pain intensity .....	70
3.4.3	Prevalence of positive responses .....	73
3.5	Discussion .....	75
3.5.1	Reliability of reports of pain reproduction .....	75
3.5.2	Reports of pain intensity .....	76
3.5.3	Prevalence of painful responses .....	78
3.6	Limitations .....	79
3.7	Conclusion and implications .....	80
Chapter 4	Measurements of Strength and Range of Movement in Painful and Non-Painful Hips .....	82
4.1	Introduction and background .....	82
4.2	Literature review .....	84
4.2.1	Reliability of measures of range of movement and strength .....	85
4.2.2	Side-to-side differences in strength in people with unilateral hip pathology .....	88
4.2.3	Comparisons of strength between symptomatic hips and normal controls .....	91
4.2.4	Side-to-side differences in range of movement .....	93
4.3	Methodological considerations .....	94
4.3.1	Instrumentation .....	94
4.3.2	Position and stabilisation of patient .....	95
4.3.3	Make versus break .....	96
4.3.4	Examiner strength .....	97
4.3.5	Duration of contraction and rest intervals between repetitions .....	97
4.3.6	Number of repetitions .....	98
4.3.7	Determining end of range .....	99
4.3.8	Summary .....	100
4.4	Methods and procedures of the current study .....	101
4.4.1	Study design, participants and procedure .....	101
4.4.2	Analysis .....	102
4.5	Results .....	104
4.5.1	Reliability of strength testing measurements .....	104
4.5.2	Reliability of ROM measurements .....	106
4.6	Discussion .....	108
4.6.1	Strength measurements .....	108
4.6.2	ROM Measurements .....	113
4.7	Limitations .....	115
4.8	Conclusions and implications .....	116
Chapter 5	The Diagnostic Accuracy of Findings from the Clinical Examination of the Painful Hip .....	118
5.1	Introduction and Background .....	118
5.2	Literature review .....	119
5.2.1	Diagnostic accuracy of physical tests .....	124
5.3	Review Summary .....	129
5.4	Methodological considerations .....	130
5.4.1	Study design & patient spectrum .....	130
5.4.2	Reference test .....	131
5.4.3	Index test description .....	141
5.4.4	Time between index test and reference standard .....	142
5.4.5	Blinding .....	142

5.4.6	Sample size analysis.....	143
5.5	Methods and procedures of the current study .....	144
5.5.1	Study design .....	144
5.5.2	Sample size.....	144
5.5.3	Participants .....	144
5.5.4	Referring specialists .....	145
5.5.5	Examiner .....	146
5.5.6	Procedures .....	146
5.5.7	Reference test .....	147
5.5.8	Index tests.....	149
5.5.9	Data analysis .....	149
5.6	Results .....	150
5.7	Discussion .....	165
5.8	Limitations .....	171
5.9	Conclusion.....	171
Chapter 6	Predictors of Intra-articular Pathology of the Hip .....	172
6.1	Introduction and background .....	172
6.2	Literature review .....	174
6.3	Methodological Considerations .....	179
6.4	Methods and procedures of the current study .....	182
6.4.1	Data collection.....	182
6.4.2	Data Analysis .....	182
6.5	Results .....	186
6.5.1	Reduced set of predictors .....	186
6.5.2	Model Selection.....	190
6.5.3	Assessment and characteristics of the model .....	194
6.5.4	Screening score versus levels of positivity .....	196
6.6	Discussion .....	199
6.7	Limitations .....	203
6.8	Conclusion.....	203
Chapter 7	The Prevalence and Diagnostic Utility of Abnormal Findings Reported in Patients Undergoing Magnetic Resonance Imaging Arthrogram of the Hip .....	204
7.1	Introduction and Background.....	204
7.1.1	Literature review .....	205
7.1.2	Prevalence of abnormal findings of intra-articular structures of the hip identified by MR imaging in asymptomatic hips .....	206
7.1.3	Prevalence of abnormal findings of intra-articular structures identified by MRI/MRA in symptomatic hips .....	208
7.1.4	Prevalence of abnormal findings at arthroscopy .....	211
7.1.5	Diagnostic accuracy of MRI/MRA .....	213
7.1.6	Review summary.....	219
7.2	Methods and procedures of the current study .....	220
7.2.1	Data Collection.....	221
7.2.2	Data Analysis .....	222
7.3	Results .....	222
7.4	Discussion .....	226
7.5	Limitations .....	229
7.6	Conclusions .....	230
Chapter 8	Summary, Key Findings and Conclusions .....	231
8.1	Key Findings .....	232

8.2	Recommendations for future research .....	241
8.3	Conclusion.....	241
	References.....	243
	Appendices .....	269
Appendix 1.	Ethical approval for interview and reliability studies .....	270
Appendix 2.	Reliability study information sheet .....	271
Appendix 3.	Consent form for reliability study .....	274
Appendix 4.	Screening and baseline questionnaire for reliability study .....	275
Appendix 5.	Consent form for interviews.....	283
Appendix 6.	Interview study information sheet.....	285
Appendix 7.	Operational definitions of tests and measures.....	288
Appendix 8.	Between-session influencing factors.....	298
Appendix 9.	Ethical approval for diagnostic accuracy studies .....	300
Appendix 10.	Screening form for diagnostic accuracy study .....	302
Appendix 11.	Diagnostic accuracy study information sheet .....	303
Appendix 12.	Consent form for diagnostic accuracy study .....	307
Appendix 13.	Medical screening questionnaire for diagnostic accuracy study.....	309
Appendix 14.	Lower limb tasks questionnaire (LLTQ).....	311
Appendix 15.	Self-report Leeds assessment of neuropathic symptoms and signs (S-LANSS).....	313
Appendix 16.	Baseline data collection form for diagnostic accuracy study .....	314
Appendix 17.	Body chart for diagnostic accuracy study .....	321
Appendix 18.	Physical examination form for diagnostic accuracy study.....	322
Appendix 19.	Percent change in pain intensity following anaesthetic injection .....	324
Appendix 20.	Coordinates of the ROC curves for mean internal ROM .....	325
Appendix 21.	Coordinates of the ROC curves for the <i>difference</i> in range of movement between painful and non-painful hips .....	326
Appendix 22.	Coordinates of the ROC curves for age .....	327
Appendix 23.	Coordinates of the ROC curves for mean BKFO ROM .....	328
Appendix 24.	Quadas 2 Critical Appraisal of DA studies for MR Imaging.....	329
Appendix 25.	MRA standardised reporting form .....	330

## List of Figures

Figure 5.1 Fluoroscopy guided anaesthetic injection .....	148
Figure 5.2 ROC curve for mean range of internal rotation of the painful hip .....	157
Figure 5.3 ROC curve for mean <i>differences</i> in range of movement between painful and non-painful hips prior to the FGAI .....	158
Figure 6.1 Best value of AICc for each predictor as a function of model size.....	191
Figure 6.2 Area under the curve versus AICc for each model with 7 or less predictors.....	193
Figure 6.3 Receiver operator characteristic curve for screening scores. ....	195

## List of Tables

Table 3.1 Prevalence of positive results for physical tests in symptomatic hips (%).....	53
Table 3.2 Frequency of reported symptoms .....	64
Table 3.3 Region of pain .....	64
Table 3.4 Pain behaviour .....	65
Table 3.5 Pain intensity (NPRS).....	65
Table 3.6 Cause, history and activity levels .....	66
Table 3.7 Baseline composite pain intensity, functional and neuropathic pain scores.....	66
Table 3.8 Within-session reliability for yes/no reports of pain reproduction.....	68
Table 3.9 Between-session reliability for yes/no reports of pain reproduction.....	69
Table 3.10 Within-session reliability for reports of pain intensity .....	71
Table 3.11 Between-session reliability for reports of pain intensity .....	72
Table 3.12 Prevalence of positive test responses.....	74
Table 4.1 Between-session mean peak force measurements for symptomatic hip (n=18).....	105
Table 4.2 Peak force measurements for asymptomatic versus symptomatic hip <sup>1</sup> (n=16).....	105
Table 4.3 Within-session ROM measurements in degrees for symptomatic hip (n=18).....	107
Table 4.4 Between-session ROM measurements in degrees for symptomatic hip (n=18).....	107
Table 4.5 ROM measurements in degrees for asymptomatic versus symptomatic hip <sup>1</sup> (n=16).....	107
Table 5.1 Overview of search terms and results per database .....	120
Table 5.2 Overview of systematic reviews of diagnostic accuracy studies investigating physical tests for intra-articular pathology of the hip .....	122
Table 5.3 Summary of quality assessment of individual studies (QUADAS-2) .....	124
Table 5.4 Baseline information.....	151
Table 5.5 Comparison between surgeon and sports physician participants .....	151
Table 5.6 Demographics and history .....	152
Table 5.7 Aggravating activities & associated symptoms .....	153
Table 5.8 Nature and area of pain.....	154
Table 5.9 Functional and neuropathic pain status .....	155
Table 5.10 Range of movement in degrees .....	156
Table 5.11 Diagnostic accuracy of ‘impingement’ tests .....	161
Table 5.12 Diagnostic accuracy of resisted movements (as pain provocation tests).....	161
Table 5.13 Diagnostic accuracy of end-range ‘rotation’ tests .....	162
Table 5.14 Diagnostic accuracy of ‘weight-bearing’ tests .....	162
Table 5.15 Diagnostic accuracy of ‘miscellaneous’ tests .....	163
Table 5.16 Diagnostic accuracy of variables from history that have a significant association with PAR (p < 0.05).....	163
Table 5.17 Post-test probability of a positive anaesthetic response .....	164
Table 6.1 Characteristics of the initial set of variables considered for inclusion in multiple logistic regression analysis ( <i>continued on next page</i> ) .....	188
Table 6.2 Summary of variables included in the reduced set of predictors.....	189
Table 6.3 Details of AICc-optimal models for each model size.....	192

Table 6.4 Details of AUC-optimal models for each model size.....	192
Table 6.5 Coefficients and odds ratios of the variables in model predicting a PAR.....	194
Table 6.6 Rescaled test coefficients (weightings) .....	196
Table 6.7 Accuracy statistics associated with rescaled screening score for predicting a PAR .....	198
Table 6.8 Accuracy statistics associated with various levels of positivity for predicting a PAR.....	198
Table 7.1 Prevalence of reported pathology PAR versus NAR (n=67).....	224
Table 7.2 Diagnostic accuracy of imaging findings (n = 67) .....	225

## Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent, has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signed:

A handwritten signature in black ink, appearing to be 'S. M. C.', written in a cursive style.

Dated: 9 August 2016

## Research Outputs Resulting from this Thesis

### Peer-Reviewed Publications

White S.G., McNair P., Laslett M., Hing W. (2015).  
Do patients undergoing physical testing report pain intensity reliably?  
*Arthritis Care Res.* 2015; 67(6):873-879.

### Submitted Publications

White S.G., McNair P., Laslett M. (2016).  
The diagnostic accuracy of symptoms and physical tests of the hip for intra-articular pathology. *American Journal of Sports Medicine*

Smythe E., White S.G. (2016) Methods of practice: listening to the story. *Physiotherapy Theory and Practice* (accepted).

### Invited Seminar Presentation

White S.G., McNair P., Laslett M., Hing W. (2012). Diagnosis of hip pain  
Nelson Hip Symposium. Nelson, New Zealand.

### Conference Presentation

White S.G., McNair P., Laslett M., Hing W. (2014).  
Do patients undergoing physical testing report pain intensity reliably?  
Physiotherapy New Zealand Biennial conference. Auckland, New Zealand

### Oral Presentations

White S.G., McNair P., Laslett M. (2015).  
Diagnostics and clinical reasoning. New Zealand Manipulative Physiotherapists  
Association Certificate in Orthopaedic Manual Therapy. Otago University,  
Dunedin, New Zealand.

White S.G., McNair P., Laslett M. (2015).  
Assessment and management of the hip. New Zealand Manipulative  
Physiotherapists Association Certificate in Orthopaedic Manual Therapy. AUT  
University, Auckland, New Zealand.

White S.G., McNair P., Laslett M. (2015).  
Assessment and management of the hip. New Zealand Manipulative  
Physiotherapists Association Certificate in Orthopaedic Manual Therapy. Hutt  
Hospital, Wellington, New Zealand.

## Acknowledgements

First and foremost, to my wife Trish, thank you for your love, support and encouragement during the period that it has taken to complete this thesis. Only you know how much time and effort it has taken and I am well aware how that has meant less quality time for you and me. You have been my rock, my gravity. You have kept me sane and healthy throughout the process by helping me to keep it in perspective. Your firm insistence that I ‘put it down’ and have a walk, or a swim, or a coffee and spend some time with friends and family was exactly what I needed. I am so looking forward to the next phase of our life together. I will pay you back for your love and patience.

To my supervisors, Professor Pete McNair and Dr Mark Laslett, thank you for your expert guidance. Having you alongside gave me the confidence to not only start this thesis, but to be sure that if I could pass your critical and experienced ‘eye’, then I wasn’t doing too bad a job! You have both been role models for me for a long time, well before our work together on this research. It has been an honour and a pleasure to have you on my team. Associate Professor Alain Vandal, you have helped to provide an extra dimension to this work. Your expertise, patience and enthusiasm are very much appreciated. Of course, Professors Liz Smythe and Wayne Hing, I haven’t forgotten you both. Along the way, you have been there to advise and encourage. It is amazing how settling it is to have someone you respect tell you that it will all be fine.

As always with a project of this nature, there are numerous people who have been involved that without whom this thesis would not have gotten off the ground, let alone be completed. Three people in particular stand out. Dr Chris Hanna, Mr Haemish Crawford and Dr Quentin Reeves. The success of this research was totally dependent on recruitment of participants and availability of resources. Chris and Haem, you guys are stars. From day one, you were on board. Not only did you diligently supply me with the patients I needed, you did this with smiles on your faces and a genuine interest in my research question. Quentin, you and your staff put up with me taking over your meeting room and rescheduling your booking systems. Your generous gift of time and expertise in performing the FGAI and MRA, and then systematically rereading all of the MR images was above and beyond the call of duty.

Of course, it goes without saying, a sincere thank you my work colleagues in the musculoskeletal team at AUT who happily took on the extra workload created by me taking time off during a busy teaching semester. That time, and your willingness to enable it, was crucial to getting this thesis completed.

To my dad Des, who passed away last year, and my mum Heather. Thank you for all you have done to help make me who I am. I have no doubt that you were both instrumental in instilling the work ethic, the self-confidence, the drive and determination that it has taken to complete this thesis. I miss you heaps Dad and wish that you were still here so that when you ask me, just once more, “how is that PhD going”, I could tell you that I have finished and that you could pat me on the back and say “well done”.

Last but not least, the rest of my family. Tanya, Luke, Lily, Max, Aaron, Emma, Nick and Ella. Thanks for understanding why we have not been able to see as much of you as we would have liked. You had better change the locks now though; I will soon have nothing to do in the weekends!

## **Ethics**

Ethical approval for the studies included in this thesis were granted by:

1. The Auckland University of Technology Ethics Committee (AUTEK) (reference number 10/44) (see Appendix 1) and
2. The New Zealand Ministry of Health, Northern X Regional Ethics Committee (Reference number NTX/11/07/066) (see Appendix 9).

Copies of the ethics approval letters and associated material such as participant information sheets and consent forms are included in the appendices.

## **Abstract**

Hip related pain places a massive economic burden on society and can have a significant effect on the individual sufferer in terms of pain, ability to participate in activity and financial costs. A delay in the accurate identification of the cause of hip pain will lead to a delay in the initiation of appropriate management, prolonging the period of suffering and possibly allowing time for further deterioration of pathology. Currently, the differential diagnosis of hip pain is based on information collected from the patient interview, a physical examination and commonly from findings identified via medical imaging. However, there is lack of good quality evidence to support the use of information from the clinical examination for either identifying or ruling out a specific cause of hip joint pain. Furthermore, there is increasing evidence that pathology identified by medical imaging is not necessarily symptomatic. The aims of this thesis were to determine the diagnostic accuracy of the clinical examination of the hip, and of MRA, for the identification of intra-articular pathology.

The thesis first explored the reliability of information gathered from physical tests of the hip. The prevalence of positive tests results and the reliability of patient reports of the reproduction of pain and of ratings of pain intensity were examined in study one. Standardised versions of physical tests were applied in a predetermined random order to both the symptomatic and asymptomatic hips of patients with unilateral hip pain. Tests were repeated one hour and 2-7 days later. The prevalence of positive findings in symptomatic hips ranged from zero to 80%, with resisted tests being the least likely to cause pain and tests that incorporated adduction or internal rotation in flexion being the most provocative. Several of these tests were also provocative in asymptomatic hips, although at a much lower prevalence. The majority of tests demonstrated 'moderate' to 'almost perfect' within-session (60 minutes apart) and between-session (2-7 days apart) reliability for both pain reproduction and ratings of pain intensity. However, the intensity of pain experienced during testing influenced reliability, with poor reliability being observed with tests that created low intensity pain. These findings indicate that clinicians can be confident that patients will reliably report both the reproduction and intensity of pain, provided the intensity is greater than 2 points on the numeric pain rating scale.

Measures of strength (peak force) obtained with a hand held force dynamometer and range of movement (ROM), obtained with a gravity dependent inclinometer, were

examined in study two. Data was collected concurrently with that of study one. Within and between-session reliability was determined and strength and ROM values between the symptomatic and asymptomatic hips were compared. Despite the presence of pain and pathology, excellent levels of reliability were observed for both peak force and ROM measures. The percent standard error of measurement for most strength tests was close to 10% and around 3% for ROM. No statistically significant differences were seen in strength between sides. However, a statistically significant reduction in range of movement was observed with the bent knee fall out test.

Study three investigated the association between information collected from a clinical examination and a positive response to a fluoroscopy-guided anaesthetic injection (FGAI). Consecutive patients with unilateral hip pain, referred for a magnetic resonance image arthrogram by selected sports or orthopaedic specialists, were recruited. Participants completed standardised questionnaires and were examined by an experienced physiotherapist, both before and immediately after the FGAI. The physiotherapist was blinded to previously obtained clinical information and the radiologist performing the FGAI was blinded to the findings of the clinical examination. A positive anaesthetic response (PAR) was defined as a reduction in pain intensity of at least 80%, calculated by comparing the mean pain intensity score from the three most provocative tests performed prior to the FGAI, to the score from the same three tests being reapplied after this procedure. A PAR was considered to indicate the presence of symptomatic intra-articular pathology. Two variables collected from the history demonstrated sufficient accuracy to indicate that they should be included in the diagnostic examination of the hip. Dominant pain in the groin had a negative likelihood ratio (LR) of 0.18 (95% CI 0.06, 0.58) and sensitivity of 0.91 (95% CI 0.77, 0.97). These values indicate that the absence of groin pain has utility as a screening test for intra-articular pathology of the hip. In contrast, the presence of crepitus had a high specificity [(0.91 (95% CI 0.77, 0.97))] and a moderate positive LR [3.67 (95% CI 1.12 to 11.9)], indicating utility for identifying intra-articular pathology. A number of physical tests demonstrated high sensitivity. However, only one of these, the quadrant test, had a negative LR [0.14 (95% CI 0.02, 1.10)] that indicated that a negative test significantly lowers the probability of a PAR. No physical test demonstrated sufficient specificity or high enough positive LR to indicate that it would be useful for identifying intra-articular pathology as stand-alone test.

In study four, a clinical prediction rule was derived from the data collected in the previous study. Logistic regression analysis was used to conduct an in-depth, systematic exploration of all possible combinations of two or more key variables using a corrected version of the Akaike information criterion (AICc) to measure model adequacy. The best overall model determined by this analysis included six variables: dominant pain in the groin; age  $\geq 39$  years; the presence of crepitus; internal ROM  $<41^\circ$ ; self-reported limited ROM and positive quadrant test. The model demonstrated an overall accuracy of 81%, sensitivity of 91%, specificity of 70%, positive and negative likelihood ratios of 3.1 and 0.12 respectively. Hence, the model has diagnostic utility to both rule in and rule out a PAR. A screening score that appropriately weights the contribution of each test finding was developed to provide a simplified means of interpreting findings for an individual patient clinically. The screening score was employed to assess the accuracy of the commonly recommended practice of ruling a specific condition in or out on the basis of the number of positive or negative tests. This demonstrated that the accuracy of this approach to decision-making is *dependent* on there being relatively equal weightings in terms of the contribution each test makes to the likelihood of a PAR. Where this is not the case, decisions made on the basis of the *number* of positive tests may be inaccurate.

Study five investigated the diagnostic utility of magnetic resonance imaging arthrogram (MRA) for identifying symptomatic intra-articular pathology of the hip, using pain response to FGAI as the reference standard. Data was collected concurrently with that obtained for study three. All participants underwent a 3 Tesla MRA immediately after the FGAI employed in study three. MR images were reported, following a standardised protocol, by a musculoskeletal radiologist with 30 years experience who was blinded to the results of the FGAI and clinical details of the participant. Despite a high prevalence of structural abnormalities identified by MRA, only the presence of subchondral bone oedema demonstrated a statistically significant association with a positive anaesthetic response. No individual structural abnormality demonstrated sufficient accuracy to indicate that it has diagnostic utility as a stand-alone finding. These observations suggest that the presence of abnormalities identified by MRA should not be considered as evidence that an intra-articular source of hip pain has been established.

This thesis provides important new information that contributes significantly to current understandings regarding the diagnostic accuracy of findings from the patient history, physical examination and MRA imaging.

# Chapter 1 Introduction

## 1.1 The problem

The diagnosis and management of hip pain places a significant cost burden on the health care system in New Zealand. In a year across 2013-2014, there were 51,832 new claims to the Accident Rehabilitation, Compensation and Insurance Corporation (ACC) for hip injuries. These claims, along with the 65,297 existing claims for on-going hip related pain, cost the ACC \$79,976,222 (Accident Compensation Corporation, 2015). As the ACC only covers the costs of injuries that result from accidents, these figures do not include costs associated with non-injury causes of hip pain e.g. osteoarthritis. In New Zealand, the prevalence of symptomatic arthritis (all joints) is projected to grow due to the demographic ageing of the population, with total financial costs of arthritis estimated to be 3.2 billion dollars a year (1.7% of Gross Domestic Product) (Access Economics, 2010). The management of painful hip pain pathology not only places a massive economic burden on society but also has significant effects on the individual in terms of pain, suffering and personal financial costs (Bennell, 2013).

Whilst a considerable amount of health care dollars are spent in the treatment of hip related pathology, accurate diagnosis of the cause of hip pain remains challenging for physiotherapists, doctors and orthopaedic surgeons (Clohisy et al., 2009; McCarthy & Busconi, 1995; Reiman, Goode, Hegedus, Cook, & Wright, 2013; Reiman, Mather, Hash, & Cook, 2014b; Tijssen, van Cingel, Willemsen, & de Visser, 2012). A recent study investigated the ability of six orthopaedic surgeons with expertise in the management of hip joint pain to make a diagnosis on the basis of detail from the patients' history and clinical tests compared to arthroscopic surgical findings (Martin et al., 2010c). This study demonstrated only 63% and 65% agreement respectively for the presence of a labral tear or femoroacetabular impingement (FAI). It has been reported that as many as 60% of patients with hip joint pain are incorrectly diagnosed on initial assessment (Byrd & Jones, 2001). Furthermore, there is evidence that some pathology in the hip can contribute to the development of other, more serious hip conditions (e.g. labral tears leading to articular delamination and degenerative joint disease, and femoroacetabular impingement leading to osteoarthritis). This suggests that more accurate and earlier diagnosis may be beneficial in slowing or stopping this progression (Ganz, Leunig, Leunig-Ganz, & Harris, 2008; McCarthy, Noble, Schuck, Wright, & Lee, 2001).

A clinical diagnosis is typically made on the basis of information collected from both the patient interview and the physical examination. Whilst several authors have suggested that information from the patient interview assists in the diagnostic process (Beattie & Nelson, 2006; Domb, Brooks, & Byrd, 2009; Jaeschke, Guyatt, & Sackett, 1994a; Peterson, Holbrook, Von Hales, Smith, & Staker, 1992; Woolf, 2003), few studies have examined the diagnostic accuracy of such information, with respect to the hip, using an appropriate reference standard (Martin & Sekiya, 2008). Similarly, there are significant issues in regard to the physical examination, including a lack of consensus, even amongst medical professionals who specialise in the management of hip joint disorders, as to which tests are most appropriate to employ, how to perform those tests and how to interpret the test results (Martin et al., 2010a; Tijssen et al., 2012). This lack of agreement may be a reflection of the lack of evidence to support the use of any particular test (or tests) for identifying a specific cause of hip joint pain (Leibold, Huijbregts, & Jensen, 2008; Rahman et al., 2013; Reiman, Goode, Cook, Holmich, & Thorborg, 2014a; Reiman et al., 2013; Tijssen et al., 2012). There is some research evidence that suggests that pain provocation tests may be useful for ruling out specific intra-articular pathology of the hip (Martin, Irrgang, & Sekiya, 2008; Maslowski et al., 2010) and that hip muscle weakness (Youdas, Mraz, Norstad, Schinke, & Hollman, 2008) and restrictions in range of motion of the hip (Altman et al., 1991; Birrell et al., 2001) are associated with hip pathology. However, there is insufficient high quality evidence to corroborate these findings.

Information gained from medical imaging is often used to complement findings from the clinical examination and advances in technology have led to the increasing use of magnetic resonance imaging (MRI), magnetic resonance imaging arthrography (MRA) and arthroscopy in the assessment of hip pain. These technologies have identified various pathologies not previously considered to be common causes of hip pain including labral tears, tears of the ligamentum teres and early articular cartilage damage (Byrd & Jones, 2004a). However, the sensitivity and specificity of MRI/MRA varies considerably depending on the structure involved, the stage of the 'disease' and the experience of the radiologist (Byrd & Jones, 2004a; McGuire, MacMahon, Byrne, Kavanagh, & Mulhall, 2012). Consequently, pathology identified by MRI/MRA may not be causing symptoms and the actual cause may not be apparent. Whilst arthroscopy can be used purely as a diagnostic procedure, there are significant costs and some significant risks associated with this procedure including infection, vascular injury and

nerve injury, making its utilisation as a diagnostic tool subject to careful consideration (Collins, Ward, & Youm, 2014; Harris et al., 2013; Kelly, Weiland, Schenker, & Philippon, 2005; Park, Yoon, Kim, & Chung, 2014).

It would be advantageous if information gathered from the clinical examination could identify pain arising from the hip without the need for such expensive and invasive diagnostic procedures (Jaeschke et al., 1994a; Leeftang, Deeks, Gatsonis, & Bossuyt, 2008; Martin, Shears, & Palmer, 2010b; McCarthy & Busconi, 1995; Tibor & Sekiya, 2008). This would provide a cost effective diagnosis, and would also be useful in areas where advanced medical imaging is not available. It may also facilitate earlier, appropriate treatment and perhaps a reduction in the incidence and severity of secondary osteoarthritis (Feddock, 2007; Groh & Herrera, 2009; Kelly et al., 2005; McCarthy et al., 2001).

Currently, no conclusive recommendations can be made regarding the value of information gathered from the clinical examination of the hip due to a lack of relevant, high quality diagnostic accuracy studies (Burgess, Rushton, Wright, & Daborn, 2011; Rahman et al., 2013; Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012). Thus, there is a need for high quality investigation of the diagnostic utility of the clinical examination of the hip so that health professionals involved in the diagnosis and treatment of people with hip joint pain can have confidence that their decisions are based on accurate and valid test findings. This thesis provides important information regarding the utility of the clinical examination of the hip based on the findings of studies that were designed to the highest possible standards.

## **1.2 Research aims**

The primary aims of the thesis were:

- To determine the diagnostic accuracy of information obtained from the patient history and from physical tests used in the clinical examination of the painful hip
- To determine if a combination of clinical tests provides improved diagnostic accuracy over stand-alone tests
- To determine the diagnostic value of ‘abnormal’ findings identified by MRA

To achieve these aims, it was first necessary to determine which clinical ‘tests’ were to be included in the central study of this thesis i.e. the diagnostic accuracy study. Both anecdotal and research evidence suggests that a large number of tests are utilised in the clinical examination of the hip. However, it was considered essential that the researcher determined that the information derived from such tests was reliable and likely to be of diagnostic value before including them in the diagnostic accuracy study.

Hence, the following questions needed to be addressed:

Question 1    What is the within and between-session intra-examiner reliability of pain provocation and of reports of pain intensity during the application of physical tests to the symptomatic and the asymptomatic hip in people with unilateral hip pain?

Question 2    What is the between-session intra-examiner reliability of measures of strength for both the symptomatic and asymptomatic hip in people with unilateral hip pain?

Question 3    What is the within and between-session intra-examiner reliability of measures of range of movement for both the symptomatic and asymptomatic hip in people with unilateral hip pain?

Once these questions were addressed, the primary aims could be focused upon and the following questions answered:

Question 4    How accurately do individual findings obtained from the clinical examination of the hip predict a positive response to an intra-articular injection of anaesthetic into the hip joint?

Question 5    What combination of findings obtained from the clinical examination of the hip best predicts a positive response to an intra-articular injection of anaesthetic into the hip joint?

Question 6    What is the prevalence of abnormal findings identified by magnetic resonance imaging arthrograms in people with a painful hip and how accurately do these findings predict a positive response to an intra-articular injection of anaesthetic into the hip joint?

### **1.3 Structure of the thesis**

Chapter 1 has introduced the problem of hip joint pain and presented the specific questions that this research addressed. Chapter 2 explores literature that has investigated the sensory innervation of intra-articular hip joint structures. This provides an understanding of which of these structures might be the source of hip pain. This chapter also provides a general review of literature relevant to the processes involved in diagnostic reasoning, the importance of diagnostic accuracy and measures of diagnostic accuracy. Thereafter, the thesis is split into chapters that specifically address the questions presented above. Each of these chapters begins with an overview and background to the related research questions. This is followed by a review of literature specific to those questions, to provide context and rationale for the study and its design.

The research itself had two key phases. Phase one (Chapters 3 and 4) used data collected from a group of participants with both a painful hip and a non-painful (normal) hip to examine various issues relevant to the application of physical tests to the hip joint. In Chapter 3, the within-session (60 minutes apart) and between-session (2-6 days apart) reliability of reports of pain provocation and of ratings of pain intensity experienced during the application of pain provocation tests was examined. The prevalence of painful responses to these tests was also determined. This study was essential given that changes in patient ratings of pain intensity following the administration of a fluoroscopy-guided anaesthetic injection (FGAI) into the hip was used as the reference standard in the subsequent diagnostic accuracy study. In Chapter 4, the intra-examiner reliability of measures of strength (peak force) and range of movement (degrees) were examined. This chapter also included comparisons of mean strength and mean range of movement between the participants' symptomatic and asymptomatic hips. This phase was an essential step in determining which tests were appropriate to include in the second phase of this research.

Phase two (Chapters 5, 6 and 7) used data collected from a group of patients with a painful hip who had been referred to a private radiology clinic by a medical specialist (orthopaedic surgeon or sports physician) for an MRA (as part of the patient's diagnostic workup). Chapter 5 examined the diagnostic accuracy of information gathered from the clinical examination of the hip by comparing those findings to the patient's response to a FGAI into the hip joint (the reference standard). Measures of accuracy including sensitivity, specificity and likelihood ratios are reported for each test. Data collected in this study was also be used to determine if any 'cluster' of

findings (i.e. a ‘clinical prediction rule’) from the clinical examination could predict the presence of hip joint pain better than findings from an individual test (Chapter 6). In Chapter 7, the prevalence of the ‘abnormal’ findings identified via the MRA of the patients included in the diagnostic accuracy study was determined and the presence of abnormal findings was correlated with the anaesthetic response for each participant. This enabled the diagnostic value of MRA findings to be determined.

#### **1.4 Significance of the research**

This research will have significance for patients with hip pain, for health professionals who treat people with hip pain and for health care funders. Improved understandings in regard to which elements of the clinical examination have diagnostic utility will enable clinicians to have more confidence that they have identified the source of a patient’s pain and subsequently in the management decisions that follow that diagnosis.

Identification of a clinical prediction rule that either ‘rules in’ or ‘rules out’ the presence of intra-articular pathology is likely to lead to a more timely diagnosis and to decrease the need for more costly and invasive medical procedures. With evidence that intra-articular pathologies such as FAI and labral tears contribute to the early onset of more serious pathologies (including chondral defects and osteoarthritis), any reduction in the delay between patient presentation and diagnosis can only be beneficial. The period of suffering for the patient and associated costs will be decreased and further deterioration of the condition may well be prevented.

## **Chapter 2 General Literature Review**

### **2.1 Introduction**

The objective of this chapter is to set the scene for the following studies by reviewing literature relevant to the clinical diagnosis of hip pain. It is complimented by additional reviews placed at the beginning of each subsequent chapter. These additional reviews provide a more focused and relevant consideration of literature that is specific to the study reported in that chapter.

This chapter begins a brief overview of the anatomy of the hip to provide context to a review of literature that has investigated the nociceptive innervation of intra-articular structures of the hip. The main study performed within this thesis explores the diagnostic accuracy of the clinical examination of the hip in regard to the identification of intra-articular pathology. In this study, people with unilateral hip pain will undergo a physical examination that establishes the intensity of any pain provoked by the application of physical tests. These tests will be repeated after the administration of a fluoroscopy-guided anaesthetic injection into the joint space of the painful hip. A significant reduction in pain intensity after this procedure will be used as the reference standard. This is based on the assumption that such pain relief indicates that a nociceptively innervated structure within the joint space has been anaesthetised and that it was likely to be the source of the patient's pain. Similarly, the absence of a significant reduction in pain intensity suggests that the pain originates from an extra-articular structure of the hip, or perhaps is referred from another body site e.g. the lumbar spine.

Next, an overview of the 'process' of diagnosis and of diagnostic reasoning is presented to give some insight into how diagnostic decisions are made. Contemporary literature relevant to diagnostics is explored and discussed so that the studies undertaken within this thesis can be viewed within the context of the diagnostic paradigm. The crucial role that information collection and interpretation plays in the diagnostic process is considered along with literature that describes the various measures of test accuracy.

## **2.2 General search strategy**

Key concepts were identified for each of the topics covered in this chapter (diagnostic reasoning, diagnostic accuracy and nociceptive innervation of intra-articular structures within the hip). Then a preliminary search was conducted, using the Auckland University of Technology (AUT) library search engine and the World Wide Web via Google, to generate an extensive list of keywords relevant to each concept. Using the keywords generated, Boolean operators (AND/OR/NOT) and truncation, a general literature search was performed via the following electronic databases: Allied and Complementary Medicine Databases (AMED), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Medline (via PubMed and EBESCO), SCOPUS and SPORT Discus (via EBESCO). The titles and abstracts were reviewed to identify and select papers relevant to the specific topic. The reference lists of the papers selected were screened to identify any other relevant papers and these were retrieved via SCOPUS. Only papers written in English were included. Further detail regarding the search strategy specific for each subsequent chapter of this thesis is presented at the beginning of the relevant chapter.

## **2.3 Intra-articular hip joint pain**

As this thesis focuses on intra-articular causes of hip joint pain and uses pain response to a FGAI as the reference standard, it is important to review the current literature in respect to the nociceptive innervation of intra-articular structures. The anaesthetic will only affect nociceptively innervated structures contained within the hip joint space. Whilst it is not the purpose of this thesis to identify precisely which structure is the source of pain, knowledge of which structures have nociceptors will enable a better understanding of how to interpret the anaesthetic response. Chapter 7 (page 204) reports the findings identified by the magnetic resonance imaging arthrogram performed immediately after the FGAI. In that chapter, associations between identified structural abnormalities and anaesthetic response were explored. The current chapter provides necessary detail that enables consideration of those findings.

The following review begins by discussing some of the broader issues surrounding musculoskeletal pain including referred pain, neuropathic pain, central and peripheral sensitisation. It is beyond the scope of this thesis to explore these issues in great depth but an overview is warranted. Following this, a brief description of the anatomy of the

hip is provided. Finally, an in-depth review of the current research into the nociceptive innervation of intra-articular structures is presented.

### **2.3.1 Pain**

Pain arising from nociceptively innervated somatic structures within the hip is commonly felt in the groin, greater trochanter or 'deep' buttock region and may also be referred to the thigh, lower leg and foot (Arnold, Keene, Blankenbaker, & DeSmet, 2011; Coomes, 1963; Leshner, Dreyfuss, Hager, Kaplan, & Furman, 2008).

It is clear that not all pain is nociceptive in nature and that neuropathic pain can result from a lesion or disease affecting the somatosensory system (either centrally or in the periphery) (Treede et al., 2008). A common manifestation of neuropathic pain is increased sensitivity to sensory input such as pressure, light touch, warmth and cold. Such sensitivity is known as central sensitisation and defined by Woolf (as cited in Lluch, 2009) as "an amplification of neural signalling within the central nervous system". Whilst increased sensitivity may indicate neuropathic pain, it is also a part of the 'normal' response to acute injury such as a sprained ankle. In this situation, local inflammation causes peripheral sensitisation of nociceptive nerve endings in the region of the damaged tissue as a part of a strategy designed to ensure that the tissue is protected from further injury. Peripheral sensitisation secondary to acute injury should not continue once the tissues damaged at the time of injury have repaired (Courtney, Kavchak, Lowry, & O'Hearn, 2010). However, with chronic musculoskeletal pathologies like osteoarthritis and chronic low back pain, persistent activity of peripheral nociceptors (resulting from the on-going chemical or mechanical irritation) causes sensitisation in the central nervous system (CNS), specifically the dorsal horn of the spinal cord (Baron, Binder, & Wasner, 2010; Courtney et al., 2010; McDougall, 2006; Thakur, Dickenson, & Baron, 2014).

The presence of central sensitisation associated with symptomatic osteoarthritis of the hip has been reported by a number of researchers (Gwilym et al., 2009; Kosek & Ordeberg, 2000; Lluch, Torres, Nijs, & Van Oosterwijck, 2014; O'Driscoll & Jayson, 1974). Some studies have demonstrated that increased sensitivity seen in patients prior to hip surgery was not present after surgery (Gwilym, Filippini, Douaud, Carr, & Tracey, 2010; Kosek & Ordeberg, 2000; O'Driscoll & Jayson, 1974). On the basis of this finding, Kosek & Ordeberg concluded that increase in pain sensitivity was due to the continued nociceptive barrage arising from the damaged structures in the hip and

was therefore not neuropathic pain. Gwilym and colleagues demonstrated that patients with symptomatic hip OA had reduced total grey matter in the thalamus (an area of the brain responsible for sensory-discriminative pain processing) compared to control subjects, suggesting neuroplastic changes in the brain as a result of chronic pain (Gwilym et al., 2010). These authors reported that the differences in volume of grey matter were only seen preoperatively and were not present 9 months after surgery, supporting the conclusions of Kosek & Ordeberg that the on-going nociceptive input into the central nervous system was responsible for the observed differences. However, the follow up period in both studies was several months after the surgery (9 months and 6-14 months respectively), a prolonged delay that would have allowed time for neuroplastic changes in the CNS to reverse. This possibility makes it difficult to determine if the on-going nociceptive barrage or neuroplastic changes in the CNS are the cause of the sensitisation (Smith et al., 2014). Regardless, these studies suggest that hypersensitivity is reversible in patients where the on-going nociceptive driver has been removed surgically.

There is some evidence that despite the presence of central sensitisation, pain relief can be achieved by blocking nociceptive input from damaged structures within a joint. Siegenthaler et al. studied patients with unilateral cervical facet joint pain and widespread hyperalgesia associated with central sensitisation and reported that a nerve block (injection of a local anaesthetic into the nerves that supply the relevant joint) completely eliminated the pain arising from that joint (Siegenthaler, Eichenberger, Schmidlin, Arendt-Nielsen, & Curatolo, 2010). This is of particular relevance to this thesis in that we used a positive anaesthetic response (>80% decrease in pain intensity reported during the application of physical tests following an injection of anaesthetic into the painful joint) as the reference standard for the diagnostic accuracy study (see Chapter Seven). The presence of central sensitisation associated with damage to somatic structures is likely to result in an increased pain response to the application of physical tests, not only to the painful structure but also to surrounding structures (Siegenthaler et al., 2010). Siegenthaler and colleagues tested pressure pain thresholds (in a manner analogous to the application of passive accessory tests by a manual therapist) in patients with pain arising from a cervical facet joint and reported that tenderness was not localised to the affected joint but instead was widespread as a result of central sensitisation. However, the elimination of pain following the nerve block indicates that

the presence of central sensitisation does not hinder the effect of anaesthetic on nociceptive pain arising from damaged tissue.

### **2.3.2 Anatomy**

It is not necessary or appropriate to provide a detailed description of the anatomy of the hip within this thesis however it is useful to define the boundaries of the intra-articular space and its contents. This section of the literature review provides a brief overview of the relevant anatomy.

The hip joint is a synovial joint comprised of the head of the femur and the acetabulum of the pelvis. It has a strong capsule that extends from the rim of the acetabulum to the neck of the femur. The capsule has a broad fibrous outer layer comprised primarily of dense collagenous tissue. This layer is lined by a thinner layer of synovial tissue that contains loose collagenous tissue, adipose tissue, a rich neurovascular network and intercellular substance (Saxler, Löer, Skumavc, Pfortner, & Hanesch, 2007). Finally, a very thin intimal layer sits on the surface of the synovial layer. Structures outside of the capsule are considered extra-articular whilst those within the capsule, bathed in synovial fluid, are considered intra-articular.

Within the capsule is the acetabular labrum, a triangular shaped fibrocartilagenous structure. The labrum is attached securely at its base to the rim of the acetabulum, except inferiorly where it is continuous with the transverse acetabular ligament that spans the acetabular notch (Freehill & Safran, 2011; Grant, Sala, & Davidovitch, 2012). Whilst in close contact with the capsule, the labrum is not attached to this structure and is separated by a small cleft. This outer aspect of the labrum is designed to resist tension and is comprised primarily of type I collagen fibres (Blankenbaker, De Smet, Keene, & Fine, 2007; Grant et al., 2012). The articular side of the labrum blends with the acetabular hyaline cartilage.

The strong ligamentum teres also sits within the hip joint space. It arises from the transverse acetabular ligament and the acetabular articular pillars and inserts into the anterosuperior aspect of the fovea capitis femoris (Cerezal et al., 2010). This ligament is composed of types I, III, IV collagen and is taut with the hip in flexion and external rotation (O'Donnell, Economopoulos, Singh, Bates, & Pritchard, 2014a; O'Donnell, Pritchard, Salas, & Singh, 2014b). The ligament is covered first with a layer of neurovascular and adipose tissue and then a layer of synovial membrane (Cerezal et al., 2010).

Finally, hyaline cartilage covers the head of the femur and the cup of the acetabulum. The acetabular fossa itself does not take part in the articulation and does not have any articular cartilage. Similarly, the region on the head of the femur where the ligamentum teres inserts is devoid of cartilage.

### **2.3.3 Innervation of intra-articular structures of the hip**

#### **The capsule**

Investigations of the macroscopic anatomy of the nerve supply to the hip joint have demonstrated that the capsule is richly innervated (Kampa, Prasthofer, Lawrence-Watt, & Pattison, 2007). A number of nerves contribute to this innervation including the sciatic, obturator, femoral and superior gluteal nerves as well as the nerve to quadratus femoris and in some people, the accessory obturator nerve (Kampa et al., 2007).

Studies that have examined the microscopic innervation of the hip capsule show that it contains both proprioceptive and pain related nerve endings (Gerhardt et al., 2012; Haversath et al., 2013; Moraes et al., 2011). Haversath et al. (2013) examined capsular, labral and ligamentum teres tissue harvested from 57 patients (mean age 55 years) undergoing elective hip surgery for a variety of conditions including osteoarthritis, avascular necrosis, FAI, developmental dysplasia and Legg-Calve-Perthes disease. In this study, samples were examined immuno-histochemically with markers that indicate nociceptive innervation (nociceptin and Substance P). Multiple slices of capsular tissue were examined so that the nociceptive innervation of the complete anterolateral aspect of the capsule could be determined. These authors reported that although the middle third of the anterolateral aspect of the capsule (close to the labral-acetabular attachment) was the most highly innervated region, the distribution of pain associated nerve fibres was relatively homogenous throughout the capsule.

Gerhart et al. used a modified gold chloride staining technique and light microscopy to identify and map the concentrations of neural receptors in specimens of capsular tissue harvested from eight cadaveric hips (mean age 76.5 years). Five of the subjects had “mild to moderate” degenerative changes and three had “severe” arthritic changes. These authors identified a “moderate number” of pain fibres in both the anterior and superolateral aspects of the capsule. However, they reported that there were not any in the inferior or posterior capsule.

Gold chloride staining and light microscopy were also used by Moraes et al. (2011) who compared tissue samples removed during total arthroplasty from patients with advanced OA of the hip (age 38-50 years) with samples removed from cadavers (age 21-50 years) with no history or evidence of OA. These authors reported that although free nerve endings (FNE) were identified in the capsule in both OA and non-OA subjects, the density of these nerve endings in patients with OA was statistically significantly reduced compared to that seen in the non-OA group. Whilst this may indicate a difference between pathological and non-pathological joints, it could be associated with differences in the age of the subjects between groups. Subsequent work by Haversath and colleagues demonstrated that increasing age was associated with a slight decrease in the population of pain related nerve fibres in the hip capsule (Haversath et al., 2013). Although Moraes et al. did not report the mean age of their subjects, the age range suggests that the mean age of those with OA was older than that of the non-OA group.

### **The synovium**

Up until the late 80's and early 90's, there was considerable debate over the question as to whether or not synovial tissue was innervated (Mapp, 1995). Whilst there is still some conflicting evidence, both sympathetic and pain related nerve fibres have been identified in synovial tissue (Mapp et al., 1990; Saxler et al., 2007; Shirai, Ohtori, Kishida, Harada, & Moriya, 2009; Takeshita et al., 2012). Mapp et al. (1990) harvested samples of synovial tissue from the fingers or knees in 5 cadavers with normal joints (mean age 39.4 years) and 5 patients having knee joint replacement surgery for treatment of rheumatoid arthritis (RA) (mean age 56.6 years). The presence of immunoreactivity to protein gene product 9.5 (PGP 9.5), substance P (SP) and/or calcitonin gene-related peptide (CGRP) was used to determine the innervation of the synovium. PGP 9.5 is a marker of sensory nerve fibres, including small diameter fibres responsible for transmitting pain (Karanth, Springall, Kuhn, Levene, & Polak, 1991), and SP and CGRP are peptides that play an important role in pain perception as well as the induction of neurogenic inflammation (Haversath et al., 2013; Takeshita et al., 2012). Mapp and colleagues reported that normal synovium contained pain nerve fibres throughout the full depth of the synovium, with some terminating in the intimal layer and others extending to the boundary between the joint space and the intimal cell layer. These authors also reported that whilst the deeper layers of rheumatoid synovium were innervated, the density of innervation was greatly reduced compared to normal

synovium. Interestingly, the superficial synovium of the rheumatoid patients was devoid of free nerve endings as were any areas of the synovium that were intensely inflamed.

In respect to normal synovium, these findings are supported by those of Saxler et al. (2007) who investigated 3 patients (aged 74 or 75 years) with painful hip osteoarthritis and compared them to 3 patients (aged 52, 64 & 80 years) with femoral neck fractures as controls. Immuno-histochemical methods were used to identify SP and CGRP nerve fibres. Pain nerve fibres were identified, primarily in the sub-intimal part of the synovial layer, in both the control patients and those with osteoarthritis. These authors reported that the density of innervation was significantly *higher* in the patients with painful OA. These findings are interesting when considered alongside those of Mapp et al. (1990) who demonstrated a *decrease* in the density of innervation in patients with rheumatoid arthritis. It is possible that these differences in innervation reflect the different joints investigated between these studies, not just the different pathologies.

In contrast to both of these studies, Shirai et al. (2009), who investigated the innervation of the synovium and labrum removed from 6 patients with OA, 3 patients with femoral head osteonecrosis and one patient with a fractured neck of femur, reported that they could not identify evidence of the presence of pain fibres in the *normal* synovium of the patients with osteonecrosis. Similarly, following a large study that investigated the inflamed synovial tissue removed during reconstructive surgery in 50 patients with symptomatic osteoarthritis and the 'normal' synovium removed from 12 patients having surgical repair after a fractured neck of femur, Takeshita et al. (2012) reported an absence of pain fibres in normal synovium. In support of the work of Saxler and colleagues, both Shirai et al. and Takeshita et al. reported evidence of pain innervation in the inflamed synovium of patients with OA. Takeshita and co-workers also performed histopathological analysis of the synovium to determine the degree of synovitis present. They reported that patients with OA demonstrated low-grade synovitis, consistent with degenerative synovitis as opposed to the high-grade synovitis seen with rheumatoid arthritis. The hip fracture patients had no evidence of synovitis. Based on these findings, these authors concluded that pain in the OA hip was associated with invasion of blood vessels and nerve fibres secondary to the inflammation of synovial tissue.

In conclusion, on the basis of the studies identified by the current search, it appears that the evidence regarding nociceptive innervation in normal synovium is contradictory.

Two small studies (Mapp et al., 1990; Saxler et al., 2007) report that normal synovium is innervated throughout its entire depth, whilst two more recent (and larger) studies (Shirai et al., 2009; Takeshita et al., 2012) reported an absence of such innervation. However, there is a consensus of evidence that demonstrates that pain nerve endings are present in the synovium of patients with symptomatic OA when there is an associated low-grade inflammation. In contrast, the synovium of patients with the high-grade synovitis associated with rheumatoid arthritis is not innervated, except for in its deepest layers where it has a sparse population of pain fibres.

### **The labrum**

Nociceptive and proprioceptive nerve fibres have also been identified within the labrum (Haversath et al., 2013; Kim & Azuma, 1995; Shirai et al., 2009). Kim and Azuma (1995) used light and electron microscopy to examine acetabular labral tissue removed from 24 'fresh' cadavers (mean age 64.8 years). These authors identified free nerve endings in all specimens. These authors did not report whether or not the study participants had a history of hip pain or any signs of pathology. However, given the broad range of subjects, it seems likely that there was a mix of normal and abnormal hips.

The previously mentioned study by Shirai et al. (2009), (see page 29 for detail) also investigated the innervation of labral tissue. In this study, immunoreactive sensory nerve fibres were identified in the labrum of patients with symptomatic OA, but only when it was associated with hyperplastic synovial tissue. This led these authors to suggest that the pain associated with OA of the hip joint was secondary to an infiltration of sensory nerve fibres into the labrum secondary to chronic synovitis. These findings contrast to those of Haversath et al. (2013) (see page 27 for study detail) who did not see any significant difference in the distribution of FNE's in the labrum despite the inclusion of patients with several different pathologies including OA, avascular necrosis and femoroacetabular impingement (FAI). Unfortunately, these authors did not report the presence or absence of associated synovial hyperplasia or inflammation making it difficult to confirm the findings of Shirai et al. Haversath and colleagues reported that the nociceptive FNE's were located predominantly at the base of the labrum (closest to its acetabular attachment) and that the density of these pain-associated nerves decreased closer to the periphery of the labrum.

### **The ligamentum teres**

The ligamentum teres (LT) has been shown to have pain nerve endings (Haversath et al., 2013; Leunig, Beck, Stauffer, Hertel, & Ganz, 2000; Sarban, Baba, Kocabey, Cengiz, & Isikan, 2007). Leunig and colleagues used light microscopy and immunohistochemical analysis to identify nociceptive fibres. Specimens were removed from 18 patients (median age 38) during hip joint surgery undertaken for a variety of reasons including femoral neck fracture, acetabular fracture, AVN secondary to slipped capital femoral epiphysis and Perthes disease. They reported the presence of Type IVa (unmyelinated free nerve endings that convey information about pain and inflammation) in all 18 subjects, regardless of the patient's diagnosis. Whilst not specifically stated by these researchers, it is likely that the LT tissue in several of the included patients was 'normal' tissue (e.g. those with femoral neck fracture) and that some was 'abnormal' tissue (e.g. those with Perthes Disease). If this is the case, this suggests that FNE are present in both the normal and abnormal LT. Leunig et al. also reported that the density of pain nerve endings was not significantly correlated with age.

Sarban et al. (2007) examined tissue from 21 children (mean age= 33.8 months, undergoing open reduction for developmental dysplasia of the hip. Whilst they identified free nerve endings in the majority of samples (58.8%), seven samples did not contain these nerve endings. This contrasts with the previous work of Leunig et al. who identified FNE's in all of their subjects. Sarban and colleagues also compared patients with partial versus full dislocation of the hip to see if the degree of tissue damage had any influence on the density of FNE's. They reported that there were not any differences between these two groups, suggesting that severity of disease does not affect the density of FNE's. These authors also reported the absence of any correlation between age and the density of FNE, supporting the earlier findings of Leunig et al., although the age range in this study was limited to 13 to 52 *months* compared to 9 to 94 *years* in the study by Leunig and colleagues.

One study (Maslon, Jozwiak, Pawlak, Modrzewski, & Grzegorzewski, 2011), has investigated the innervation of the LT in 19 patients with cerebral palsy undergoing open reduction of a dislocated hip. This study demonstrated a significant ( $p= 0.0001$ ) correlation between the density of pain nerve endings and the pain intensity reported by the participants (preoperatively). Higher density of pain nerve endings was associated with higher pain intensity. Similarly, the density increased in those with loss of cartilage and this increase was more significant as the degree of cartilage damage increased.

The most recent study (Haversath et al., 2013) identified in the current search was described previously with respect to innervation of the capsule and labrum. These authors also examined the LT and reported that pain nerve endings were present throughout the LT. The highest concentration of pain nerve endings was seen in the centre of the ligament, closely associated with the blood vessels. Density was lower at both the acetabular and femoral ends of the ligament. All patients in this study were undergoing surgery for painful hip pathology (including OA, AVN and FAI).

In summary, apart from Sarban et al., who reported the absence of pain nerve endings in *some* of the young children having surgery for developmental dysplasia of the hip, there is a consensus of evidence that supports the presence of pain nerve endings in the ligamentum teres. Whilst this evidence suggests that the population of pain nerve endings in this structure does not change as a result of aging, the population increases when there is associated cartilage damage. Furthermore, increasing degrees of cartilage damage appear to lead to higher concentrations of these nerve endings.

#### **Articular cartilage and subchondral bone**

No studies that have explored the innervation of articular cartilage in the hip joint were identified in the current search. There appears to be a general consensus that *normal* articular cartilage does not contain nerves, blood or lymph vessels (Maslon et al., 2011; van Dijk, Reilingh, Zengerink, & van Bergen, 2010). There is some evidence that there are sensory nerve endings in the deepest layer of articular cartilage but there is little to suggest that these fibres come into contact with chondrocytes in the growth plate or outer layers of the cartilage in humans (Grassel, 2014). However, there is some evidence that damaged articular cartilage can contain both blood and neural tissue (Ogino et al., 2009; Suri et al., 2007; Szadek, Hoogland, Zuurmond, De Lange, & Perez, 2010). The earliest of these studies (Suri et al., 2007) identified pain nerve fibres associated with blood vessel ingrowth into the articular cartilage from the underlying subchondral bone. Participants in this study were primarily patients undergoing total knee joint replacement as a treatment for symptomatic OA. Nerve fibres immunoreactive to PGP 9.5, SP and CGRP were observed at the osteo-chondral junction as well as within marrow cavities of osteophytes and within the subchondral bone marrow. Suri and colleagues also reported that these findings were seen across patients with a wide range of severity of OA, not just in the end-stage patients. Ogino et al. investigated the innervation of sub-chondral bone in 15 patients undergoing total knee arthroplasty (TLA) for severe medial compartment OA. These authors compared

sections from the affected medial component to sections from the normal lateral compartment from the same patient. Ogino et al. reported the presence of substance P, Cox-2, TNF- $\alpha$  and TUJ1 in the subchondral plate of the affected compartment but no evidence of innervation of the subchondral bone in the normal compartment.

Szadek et al. (2010) investigated the innervation of articular cartilage of the sacroiliac joint in 10 human cadavers (mean age 69.8) using immunoreactivity to substance P and CGRP as indicators of nociceptive innervation. These authors reported immunoreactivity to both SP and CGRP in the superficial layer of both the sacral and iliac cartilage in nine of the 10 cadavers. The authors noted the presence of degenerative changes of the articular cartilage and commented that the immunoreactivity that they observed might represent extracellular deposits of SP and CGRP secondary to this degeneration. Unfortunately, it could not be established if the patients had any history of symptoms related to the SIJ that might suggest that these changes were relevant. Sensory nerves were not identified in subchondral bone.

It is difficult to compare the findings of Szadek et al. to the two previous studies given that they were conducted in patients with OA severe enough to warrant total knee joint replacement whereas Szadek and colleagues investigated the sacroiliac joints of cadavers that may well have been asymptomatic. However, all of these studies point towards the fact that articular cartilage in joints with degenerative changes may well be innervated by pain nerve endings. These findings may not be relevant to hip articular cartilage.

#### **2.3.4 Review Summary**

This literature provides evidence that the capsule, the synovium, the labrum, the ligamentum teres and articular cartilage may contribute to pain arising from within the hip. Whilst the capsule is richly innervated, particularly in its anterolateral aspect, the population of nerve endings appears to reduce in the presence of symptomatic OA. One small study, performed on elderly cadavers, has reported that the inferior or posterior capsule is not innervated. The evidence regarding innervation of the normal synovium is conflicting. However, most authors have reported innervation of the synovium in the presence of inflammation secondary to OA.

Sensory nerve fibres are present in the labrum, with a higher density at the base than the periphery. Some evidence exists to suggest that this innervation is associated with hyperplastic synovial tissue in patients with symptomatic OA. Similarly, the

ligamentum teres is normally innervated. Interestingly, the density of pain nerve endings in this ligament appears to be higher in patients with higher pain. There is some contradictory evidence regarding a correlation between the severity of associated cartilage damage and the density of pain nerve endings. Normal articular cartilage is not innervated although some research has demonstrated that it may contain sensory nerves in people with osteoarthritic joints secondary to vascular ingrowth.

## **2.4 The diagnostic process**

The process of collecting information from the history and clinical examination in an attempt to determine the cause of a patient's symptoms and signs has long been the mainstay of medical practice. The word diagnosis is derived from the Latin word *diagignōskein* meaning "to distinguish". *Diagignōskein* is derived from the Greek words *dia* meaning "apart" and *gignōskein* meaning "to learn" or "to know" (Harper, 2015). Whilst the making of a diagnosis was clearly a part of the process in managing patients in the ancient Greek schools of medicine at the time of Hippocrates (c 460-370 BC), there is evidence that the Egyptian physician Imhotep (c 3000-2500 BC) made a diagnosis (that defined both the treatment and prognosis of a patient) on the basis of information collected from the history and objective examination (Brandt-Rauf & Brandt-Rauf, 1987).

There has been a considerable amount of research and debate about the cognitive processes involved in diagnostic reasoning in clinical medicine over the last 40 years (Elstein, 2009; Elstein & Schwarz, 2002; Monteiro & Norman, 2013). Research has been dominated by consideration of two key theories i.e. hypothetico-deductive reasoning and pattern recognition. The 'hypothetico-deductive' theory, the earliest model of reasoning, is based upon the formulation of hypotheses (with an order of most to least likely), testing of these hypotheses (via examination of data collected from the patient examination) and subsequent modification or re-ranking of the hypotheses until the clinician is confident that the most likely cause of the patient's problem has been identified. Data collection is guided by the initial hypotheses and modification and re-ranking occurs when additional findings either support or place in doubt the primary hypothesis. This process is analytical and time consuming. It is based on the Bayesian method of considering the probability of a particular diagnosis being present (based on the clinician's own experience or perhaps by published evidence) prior to the collection of new data (e.g. results of a diagnostic test or the patient's answer to a new question from the clinician) and then again after consideration of that new data.

In this model, poor initial hypothesis generation (e.g. clinicians with an inadequate knowledge base or experience), inappropriate data collection (e.g. using a test that is unreliable or not valid) and/or misinterpretation of the data collected (e.g. inaccurate reading/measurement, misunderstanding between patient and clinician) can have a significant effect on the diagnostic decision (Elstein & Schwarz, 2002). Apart from this type of error, the method is also subject to a number of other factors that can significantly influence decision-making including fatigue, inattentiveness, distraction and cognitive overload. Research in this area has demonstrated that both novices and experts use hypothesis testing and that it appears that it is the depth of knowledge of the clinician that is the primary determinant of a successful outcome (correct diagnosis) rather than the process of generating and testing of hypotheses (Elstein, 2009; Elstein & Schwarz, 2002; McLaughlin, Eva, & Norman, 2014).

In contrast to the hypothetico-deductive model, the pattern-recognition model is intuitive and rapid. In this model, it is proposed that experienced clinicians recognise overall patterns in patient presentation almost instinctively. The clinician employs mental shortcuts (aka heuristics) such as ‘common sense’, ‘rules of thumb’ or ‘I’ve seen this before’ to make an ‘educated guess’ at the diagnosis. In this model, diagnostic accuracy is dependent on the clinician’s knowledge base rather than the process of decision-making (Croskerry, 2009; Elstein, 2009; Elstein & Schwarz, 2002). There is evidence that ‘Experts’ often forgo explicit hypothesis testing having faith that their extensive experience and mastery of the knowledge base can be relied upon (Norman, Coblenz, Brooks, & Babcook, 1992). This method of decision-making is consequently dependent on the depth and breadth of the clinicians knowledge base and their ability to be able to retrieve, from their memory, ‘packages of information’ that represent an accurate clinical diagnosis. Whilst more clinical experience provides more opportunity to recognise and learn different patterns of symptoms and signs (representing unique diagnostic categories), it does not ensure success in the diagnostic process. Heuristics are subject to various cognitive biases including framing, anchoring, overconfidence and omission (Croskerry, Singhal, & Mamede, 2013a). A lack of awareness that these biases are likely to be present, or a lack of a willingness to institute strategies to address such biases, is likely to have an adverse affect on the reasoning process and ultimately the conclusions made by the clinician (Croskerry, Singhal, & Mamede, 2013b; Elstein, 2009).

Dual-process theory proposes that the hypothetical-deductive and pattern recognition/intuitive models of diagnostic reasoning are not necessarily dichotomous but instead that they sit along a continuum (Croskerry, 2009; Monteiro & Norman, 2013). Croskerry proposed a model of diagnostic reasoning, based on dual-process theory, that provides a framework that helps explain the interactions between the intuitive approach, which he calls 'System 1' processes and analytical approach ('System 2') (Croskerry, 2009). Whilst dual processing models appear to dominate current research in this area, it is important to note that there are numerous variations of this basic theory. No one universal model has been agreed upon, primarily because none has been identified that explain all of the characteristics of diagnostic decision making that have been observed amongst clinicians with various levels of experience and/or across various clinical situations (Monteiro & Norman, 2013).

What does seem to be a consistent observation in this literature is that the content knowledge of clinicians is a key factor in their ability to make accurate clinical diagnoses (Elstein, 2009; McLaughlin et al., 2014; Monteiro & Norman, 2013). Whether it be knowledge gained through clinical experience (and reflection), through implicit learning (case studies, course, conferences) or study of research evidence, and regardless of whether it be knowledge that feeds the intuitive or analytical systems of reasoning, the larger the knowledge base, the more likely an accurate diagnostic decision will be made.

The obvious corollary of this understanding is that this content knowledge needs to be sound. Decisions made on the basis of incorrect information are less likely to lead to accurate diagnoses and consequently may lead to ineffective, or even inappropriate, treatment. Gawande (as cited in Croskerry, 2009) reported that autopsy findings demonstrate that as many as 40% of diagnoses made when the patient was alive are incorrect and frighteningly, that if a correct diagnosis had been made (and appropriate treatment administered), then one third of patients being autopsied would not have died.

Whilst the consequences of poor diagnostic decision making in regard to the management of hip pain are unlikely to lead to death, serious consequences may result from a failure to identify pathologies such as infection, osteonecrosis or a slipped upper femoral epiphysis (Assouline-Dayana, Chang, Greenspan, Shoenfeld, & Gershwin, 2002; Maj, Gombar, & Morrison, 2013). Even with more benign causes of hip pain such as femoroacetabular impingement (FAI), there is evidence that delayed treatment

secondary to lack of an early accurate diagnosis may lead to long term consequences (Ganz et al., 2008; Leunig, Beaulé, & Ganz, 2009). Similarly, ‘over-diagnosis’ resulting from inappropriate testing or over-emphasis on some test findings can have significant effects on management and outcomes (Deyo, 2013; Larson et al., 2015; Tang, 2007).

Information collected from history and physical examination provide the basis for the making of a diagnosis (Feddock, 2007; Hampton, Harrison, Mitchell, Prichard, & Seymour, 1975; Peterson et al., 1992; Woolf, 2003). Rather than make a decision on the basis of just one finding, clinicians consider the overall pattern or ‘picture’ of the patient before them (Feddock, 2007). However, information collected during the diagnostic process influences subsequent diagnostic decisions. In theory, each test finding should increase or decrease the probability of a given diagnosis being present (Jaeschke et al., 1994a). However, if data collected early during the assessment is inaccurate or misleading, subsequent assessment may be inappropriately focussed. Consequently, the accuracy of such information is imperative. As few tests are 100% accurate, knowledge of the level of accuracy of individual findings from the clinical examination of a patient allows the clinician to consider the value of that finding for that patient.

Whilst a number of researchers have explored the reliability and diagnostic accuracy of data collected from both the history and clinical examination of patients with hip pain of musculoskeletal origin, most commentators agree that there is insufficient, high quality evidence to support the use of the majority of tests currently utilised in this area of practice (see Chapter 5 for an in-depth review of this literature) (Burgess et al., 2011; Leibold et al., 2008; Rahman et al., 2013; Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012).

## **2.5 Performance characteristics of tests**

A reliable test will give consistent results with repeated measures. It will have only small errors in measurement and hence will allow the clinician to use information gained from that test with confidence (Haas, 1991; Jensen, Wang, Potts, & Gould, 2012). Whilst it is essential for diagnostic tests to be reliable, they must also be valid (Delitto & Snyder-Mackler, 1995). A valid test will accurately measure what it is supposed to measure. The diagnostic accuracy of a test is a measure of how valid that particular test is at correctly identifying or ruling out a specific pathology. If the reliability and accuracy of the test is unknown, then diagnostic and management decisions may well be based on incorrect information.

There are a number of metrics that are used to describe diagnostic accuracy. The following review provides an overview of those most commonly used. It describes the different aspects of accuracy that each metric reflects alongside a discussion regarding its key strengths and weaknesses. This will provide an understanding that will enable the most appropriate decisions to be made in regard to which test to employ given the circumstances and intent of its use (e.g. to identify or to screen out a specific pathology). This has particular relevance to the diagnostic accuracy study performed in Chapter 5. More focused reviews of the literature concerning the reliability of measures of pain, strength and ROM and of the diagnostic accuracy of physical tests for the hip joint are presented in subsequent chapters of this thesis.

### **2.5.1 Sensitivity and specificity**

Perhaps the two most commonly reported measures of test accuracy are sensitivity and specificity. These measures provide detail about the likelihood that a *test finding* will be accurate. A sensitive test will be positive in people known to have the condition of interest and a specific test will be negative in those known not to have the condition. The level of sensitivity is determined by calculating the proportion of patients with the disease in whom the test is positive. Similarly, the level of specificity is determined by calculating the proportion of patients without the condition in whom the test is negative.

A test that is always positive in people with a particular condition is 100% sensitive. With such a test, a negative result can be relied upon to ‘rule out’ that condition, given that there will not be any false negatives with this test. Similarly, a test that is always negative in people without the condition is 100% specific. With this test, there will never be any false positives, enabling the clinician to ‘rule in’ the condition in the event of a positive test result. Sackett (1992) proposed the use of the mnemonics SnNout (where *Sensitivity* is high and the test result is *Negative*, rule the condition *out*) and SpPin (where *Specificity* is high and the test result is *Positive*, rule the condition *in*). However, few tests are 100% accurate and the possibility of false test results is usually present. Pewsner et al. (2004, p. 212) cautioned “...the power to rule a disease in or out is eroded when highly specific tests are not sufficiently sensitive, or highly sensitive tests are not sufficiently specific”.

Knowledge of the level of sensitivity and specificity of a test gives an estimate of the likelihood that the test result will be a true reflection of the status of the condition e.g. with a sensitivity of 80%, there is an 80% probability that the test will be positive in

someone with the condition and a probability of 20% that the test will be a false negative (Griner, Mayewski, Mushlin, & Greenland, 1981). There is no apparent consensus as to what level of sensitivity or specificity is acceptable for clinical use, in part because such levels are likely to change depending on the severity of the consequences of making the wrong diagnostic decision (Griner et al., 1981; Haneline, 2007; Sackett, 1992). For example, a sensitivity of 80% is probably acceptable for a test that is being used to rule out a gluteal tendinopathy, whereas this level of sensitivity might be considered inappropriate to rule out a slipped upper femoral epiphysis in a 13 year old with a painful hip.

Whilst an understanding of the probability of a true or false test result is useful, this does little to inform the clinician about the probability of the condition of interest being present (Aliu & Chung, 2012; Griner et al., 1981). In the clinical situation, the status of the condition is not known. During the diagnostic process, the clinician is trying to determine which condition is most likely to be causing the patient's symptoms/signs. Instead of asking what is the likelihood of a test result being accurate, the clinician wants to know what is the probability that a condition is present given the results of particular test.

### **2.5.2 Predictive values**

This question is answered, to a degree, by calculating the predictive value of the test. The positive predictive value (PPV) is the proportion of people who test positive that have the condition (i.e. true positives) and represents the probability that the condition will be present (Griner et al., 1981). Conversely, the negative predictive value (NPV) is the proportion of people who test negative that do not have the disorder (true negatives) and represents the probability of the condition being absent. A PPV of 75% informs the clinician that, in the event of a positive test finding, there is a 75% probability that the condition will be present. Similarly, with a NPV of say 95%, there is a 95% probability that the condition will be absent when that test is negative.

A limitation regarding the use of predictive values, compared to sensitivity and specificity, is that they are influenced by the prevalence of the condition in the cohort of patients in which tests are being used or evaluated (Sackett, 1992). The prevalence of a condition will vary depending on a number of factors including, but not limited to, the age of participants (e.g. degenerative conditions are more likely with increasing age), the reference standard used to determine the presence of the condition (e.g. labral tears

identified at surgery versus MRA or MRI), the definition of a positive index test (e.g. reproduction of pain versus loss of ROM) and the definition of a positive reference test (e.g. a 50% reduction in pain after an intra-articular anaesthetic injection compared to a 80% reduction) (Griner et al., 1981).

As an example, the FABER test was investigated by Maslowski et al. (2010) in 50 patients, aged between 22 and 76 years (mean 60 years), who had an intra-articular anaesthetic injection of their hip as the reference standard. Twenty-six patients reported at least an 80% reduction in pain intensity after this procedure and were considered to have symptomatic intra-articular pathology. These authors reported that the FABER test had sensitivity of 81% (95% CI 58-95) and specificity of 25% (95% CI 8-50). Based on the response to the anaesthetic injection, the prevalence of the condition of interest in this study was 52%. The PPV was reported as 54% (95% CI 35-72) and the NPV as 55% (95% CI 20-86). Other authors (Martin et al., 2008) have used a 50% reduction in pain intensity as the reference standard. If Maslowski et al. had used a 50% cut-off point, the prevalence of intra-articular pathology in their study would have increased to 70%. Sensitivity and specificity would have remained similar, at 80% and 26% respectively. However, the PPV would have increased to 71% and the NPV would have dropped to 36%. Conversely, if a 100% reduction in pain intensity after the anaesthetic had been considered as the reference standard, the prevalence of intra-articular pathology would decrease to 30%, sensitivity would increase slightly to 87%, specificity would stay at 26%, but the PPV would drop to 33% and the NPV would rise to 81%.

For predictive values reported for a given test to be useful in another setting, users of the test need to ensure that the prevalence of the condition of interest in the new setting is similar to that in the study in which the test was evaluated. Fortunately, this is not an onerous or additional task in that the rationale for applying any test should only be to confirm or deny the presence of a condition that is likely to be present. Thus, the clinician already has a suspicion that the condition is present, based on the patient characteristics, history, symptoms and signs. This is discussed in more detail in regard to likelihood ratios.

### **2.5.3 Likelihood ratios**

Likelihood ratios (LRs) are an alternative measure of test accuracy that provide a number of advantages over sensitivity, specificity and predictive values. Likelihood

ratios combine the information contained in both sensitivity and specificity and express the *odds of a given test result* occurring in a patient with the condition of interest compared to without that condition (Davidson, 2002; Sackett, 1992). Both positive and negative likelihood ratios are often reported in diagnostic accuracy studies. Positive LR (LR+) reflect the change in odds of a condition being present based on a positive test result whereas a negative LR (LR-) reflects the change in odds that the condition will be present when that test result is negative (Aliu & Chung, 2012). Sackett (1992, p. 2643) argues that LR's are "much faster and more powerful than the sensitivity and specificity approach". Unlike predictive values, LR's are not influenced by prevalence (Denegar & Fraser, 2006; Simel, Samsa, & Matchar, 1991).

LR's are relatively easy to interpret. LR's greater than 1.0 indicate that the probability of a specific condition being present has increased as a result of that test finding. The larger this value, the more probable (higher odds) that the condition will be present. For example, a LR of 15 reflects that a *positive test result* is 15 times as likely to occur in a patient with the condition than in one without that condition (Jaeschke et al., 1994a; Sackett, 1992). LR's less than 1 indicate that, in the event of a negative test result, the probability of the condition of interest being present has decreased. The smaller the negative value, the smaller the probability (lower odds) that the condition will be present and the more useful a negative test result is for helping to rule out that condition (Sackett, 1992). LR's of 1 indicate that the pre-test and post-test probability of the presence of a specific condition is exactly the same after the test as it was before the test i.e. the test has no diagnostic value.

As discussed with predictive values, levels of accuracy for a given test can change depending on the circumstances in which they are used. For predictive values, the prevalence of the condition of interest is a significant factor. The severity of the condition in the cohort in which a test is being utilised is another factor that will influence test accuracy (Jaeschke et al., 1994a; Pewsner et al., 2004). The more severe the condition, the more sensitive a test is likely to be. With an increase in sensitivity, both positive and negative LR's shift further from the value of 1 reflecting an improvement in the ability of the test to distinguish between those with and those without the condition. Hence, tests evaluated in patients that have undergone surgery (suggesting that they have a condition severe enough to warrant surgical intervention) will demonstrate better accuracy than those evaluated in a primary health care setting. Expecting the levels of accuracy across these two settings to be the same is unrealistic.

Bearing this in mind, an advantage of LRs is that they provide a means to quantify the actual odds of a *condition being present*, as opposed to the probability of a positive or negative *test result* occurring in someone with the condition (Jaeschke et al., 1994a). To do this, the clinician must first make an estimate of the probability of the condition being present in the patient before they apply the test. Such an estimate may be based on evidence from the literature regarding the prevalence of that condition in specific populations, or, more commonly, based on the clinician's clinical experience. For example, an experienced clinician might estimate that the likelihood that a female, aged 18 years, who injured her knee playing netball (by landing suddenly in a manner that forced knee hyper-extension with internal rotation) and who reported immediate pain and significant swelling, has an 80% chance of having damaged her anterior cruciate ligament (ACL). Next, this pre-test probability is converted into pre-test odds (using the formula  $\text{odds} = \text{probability} / (1 - \text{probability})$ ). In this example, the pre-test odds is 4, meaning that the netballer is 4 times more likely to have damaged her ACL than someone without this history.

Based on this evidence, it would be appropriate for the clinician to investigate the integrity of the ACL using a test with a high level of diagnostic accuracy. One such test is the prone Lachman test, which has a positive likelihood ratio of 20 (Mulligan, Harwell, & Robertson, 2011). In the event of a positive test, the post-test odds of an ACL injury is calculated by multiplying the pre-test odds by the positive LR. Using the current example, the odds of this netballer having an ACL injury has now risen to 80% ( $4 \times 20$ ). Finally, the post-test odds is converted back to probability ( $\text{post-test odds} = \text{post-test odds} / (\text{post-test odds} + 1)$ ). In the current example, the probability of an ACL tear has increased to 99% as a result of this positive test finding. Whilst this process is cumbersome, post-test probabilities can be obtained easily through the use of nomogram proposed by Fagan (Fagan, 1975), provided the LR of the test is known and the clinician has estimated the pre-test probability of the condition being present.

Another advantage of LRs is that they provide a means to consider the multiplicative effect of more than one test result in that the post-test probability of a condition being present after one test becomes the pre-test probability for the subsequent test. Each item of history & physical examination can be considered a diagnostic test in its own right and can either increase or decrease the probability of the condition of interest being present (Jaeschke et al., 1994a). However, this is only the case when there is not too close a relationship between the included tests, i.e. each test needs to provide

information that is distinctly different and ‘independent’ (Jaeschke et al., 1994a; Sackett, 1992).

#### **2.5.4 Overall accuracy**

Each of the abovementioned metrics provides different and specific information regarding the diagnostic accuracy of a given test. Where these values are known and the purpose of further testing is clear (e.g. a highly sensitive test is needed to rule out a competing diagnosis), it is often not difficult for clinicians to select the most appropriate test to use. However, sometimes the relative merits of one test over another are not so obvious and it would be useful if there were a single indicator that provided an overall measure of the accuracy of a test to guide test selection. A number of statistics have been proposed as a way of summarising test performance, the most common of which are discussed below.

##### **Youden index**

The Youden index (YI) is calculated by subtracting 1 from the sum of sensitivity and specificity) and represents the total correctly classified rate (Youden, 1950). A value of 1 indicates that there are no false positives or false negatives. A value of 0 indicates that the same proportion of positive test results was seen in people with and those without the condition i.e. the test has no discriminatory power. The major disadvantage with this index is that two tests can have the same YI despite having very different sensitivity and specificity. Consequently, it should always be considered in conjunction with these other metrics.

##### **Accuracy index**

Another statistic used to give an indication of overall test performance is the ‘accuracy index’ (aka ‘accuracy’). The accuracy index (AI) is simply the percentage of individuals tested with correct results (true positives and true negatives). Aliu and Chung (2012) suggest that this measure is not particularly useful in that it does not distinguish false positives from false negatives, a distinction that has importance clinically. When sensitivity and specificity are not equal, the AI is affected by the prevalence of the target condition in the cohort in which the test is being examined (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003). Thus, the reported AI of a given test may not reflect how that test will perform under different circumstances.

### **Diagnostic odds ratio**

A third alternative for giving an indication of the overall performance of a specific test is the diagnostic odds ratio (DOR). The DOR is the ratio of the odds of positivity in people with the condition to the odds of positivity in those without the condition (Glas et al., 2003). Thus, the DOR is a measure of strength of the association between the test finding and the condition. It is not affected by prevalence of the condition of interest. Whilst there are a number of ways to calculate the DOR, perhaps the most simple means is to divide the LR+ by the LR- (Glas et al., 2003). A test with a DOR of 1 has no discriminatory power i.e. it will not be able to distinguish people with a condition from those without that condition. As test performance improves, the DOR value increases (with no upper limit). A major advantage of the DOR is that it is easy to interpret e.g. for a test with a DOR of 52, the odds of a positive test result in someone with the condition of interest is 52 times higher than the odds of a positive test in someone without that disorder. Thus, comparison of the overall diagnostic strength of two or more tests is simple.

As LRs are derived from sensitivity and specificity, the information contained in these metrics is incorporated into the DOR. However, like the Youden index, two tests with the same DOR may have very different sensitivity and specificity. Hence, in a clinical situation where it is important to know the error rates, sensitivity and specificity would still be the metric of choice. For example, if screening to rule out a serious condition (say avascular necrosis or infection), knowledge of the likelihood of a false negative result associated with a test is crucial and would require the clinician to consider the test sensitivity. As with all other measures of test performance, the DOR is affected by the severity of the condition and the criteria for defining a positive test results (both index and reference tests).

Glas et al. (2003) argue that the DOR is a measure worth utilising when the combined diagnostic value of a number of test findings is being explored such as when using logistic regression analysis to construct clinical prediction rules (CPR). This type of statistical analysis determines the degree of accuracy of various combinations of test findings, taking into account any interactive effects of one test result on another (Guyatt et al., 1995). The DOR for various combinations of test findings can be compared to help identify the CPR that has the optimum clinical utility.

### Receiver operator characteristic curves

Range of movement, muscle strength and participant age are examples of information gathered from the history or clinical examination that have a continuous rather than a dichotomous outcome. When these findings are compared to a reference standard, the cut-off point selected to represent a positive test, will to a large degree, determine the level of accuracy for that test. Receiver operator characteristic (ROC) curves provide a way of comparing sensitivity and specificity at any cut-off point. They are constructed by plotting the true positive rate (sensitivity) of the test against the false positive rate (1-specificity) for that test at each cut-off point (Hagen, 1995). This produces a visual representation of the trade off between sensitivity and specificity depending on the cut-off point selected.

The area under the resulting 'curve' on this graph is another measure that represents the overall performance of a test. It can be interpreted as the average sensitivity over the entire range of specificities or vice versa (Eng, 2005). A test that has 100% sensitivity and specificity will have an area under the curve (AUC) of 1. A test that has no diagnostic value would have an AUC of 0.5 and the 'curve' on the graph would in fact be a straight line running diagonally from the bottom left to the top right of the graph. AUC values less than 0.5 do not necessarily mean that the test has no diagnostic value. In fact, the closer the value approximates zero, the more test accuracy improves *provided that* the test result is inverted. For example, it might be expected that age *over* 50 would be a predictor of hip osteoarthritis. If this association is explored in a diagnostic accuracy study and the AUC for this age at this cut-off point is very low (say 0.12), this indicates that age *over* 50 is a poor predictor of OA. However, such a small AUC indicates that age *under* 50 is in fact a good predictor of hip OA.

An advantage of ROC analysis is that it gives an opportunity to consider the test accuracy over all possible cut-off points and prevents the potential loss of information that might result if just one, predetermined, point is used to define a positive test. Cut-off points can be selected to suit the intended use of the test, for example, if high sensitivity is required so that a test can be used to rule out a specific condition, the point closest to the upper right portion of the curve should be selected (Griner et al., 1981; Hagen, 1995). False positives are of concern with a test that is used to identify a condition that might require a subsequent invasive procedure (e.g. surgery) (Hagen, 1995). In this case a cut-off point closest to the lower left portion of the ROC curve will

minimise the number of false positives and therefore the risk associated with exposure to such procedures.

### **2.5.5 Review Summary**

There is usually a degree of uncertainty during the diagnostic decision making process. Whilst numerous indicators of test performance are available, none provide a stand-alone measure that allows the user of a diagnostic test to attain a complete picture of the degree of uncertainty associated with the use of that test. Indeed, many of the metrics used can be misleading given the fact that most are influenced by characteristics of the population of interest as well as the methods employed in the studies that have evaluated the test performance. Recognition of this factor is crucial to understanding diagnostics and the appropriate selection of clinical tests and the interpretation of their findings.

Ideally, clinicians should consider the similarities between the patient to whom they intend applying a given test and those included in the study that evaluated that test. Tests need to be performed and interpreted in the same manner, in a patient that has similar characteristics and a similar stage of disease as those in the diagnostic accuracy study. To quote Jaeschke et al. (1994a):

If you practice in a setting similar to that of the investigation and your patient meets all of the study inclusion criteria and does not violate any of the exclusion criteria, you can be confident that the results are applicable. If not, a judgement is required. (p. 706)

Given a situation where the clinician decides that there is sufficient similarity to justify the use of a test, the findings of the test are best interpreted with a combination of test metrics. Ruling in or out a condition on the basis of one test could be considered inappropriate. Doing so on the basis of one measure of accuracy of a single test might be considered foolhardy.

Studies that investigate the diagnostic accuracy of tests should be designed in a manner that optimises the generalisability of their findings. The tests being evaluated should have proven reliability and researchers should provide a detailed description of how to perform and interpret each test. This will enable others to replicate the test and to have confidence that they will make the same decisions regarding the findings of that test. Diagnostic tests should only be used in clinical practice when there is a sufficient degree of suspicion that a given condition is likely to be present. Hence a broad

spectrum of participants should be included in diagnostic accuracy studies so that it is representative of the population of patients to which the test is most likely to be used in clinical practice. Study participants should have symptoms and/or signs that suggest the presence of the condition of interest. The stage and severity of the condition of interest should be as wide as possible i.e. not just end stage disease that is easily identifiable. Finally, researchers should report study findings using a wide range of test metrics so that clinicians can get a more complete understanding of the tests diagnostic utility. Preferably, raw data (the number of true and false positives and negatives) should be reported so that users of the research can do their own analysis of test performance.

## **2.6 Chapter summary**

This chapter has discussed key issues relevant to the diagnosis of intra-articular pathology of the hip joint. It has provided an up-to-date review of the sensory innervation of the intra-articular structures of the hip joint space. It is clear from this evidence that all intra-articular structures are capable of causing pain. This is particularly likely in joints with pathological changes that alter the innervation of tissue e.g. osteoarthritis leading to innervation of articular cartilage (which is not innervated in its normal state) and to an increase in density of nerve endings and fibres in the synovium and labrum. This information will enable both clinicians and researchers to interpret responses to intra-articular injections of anaesthetic.

The chapter also considered the process of diagnosis and demonstrated that this process is essentially one based on probability where, given the patients presenting symptoms and signs, the clinician must first estimate the probability of the presence of the various conditions capable of causing the patient's problem. Based on these estimates, further evidence that either strengthens the probability of the presence of the most likely diagnosis, or decreases the probability of the presence of a competing diagnosis is obtained and considered. Hence, the collection and correct interpretation of additional evidence from the clinical examination is crucial. The choice of test to employ and how to interpret the findings of that test should only be made on the basis of the test metrics. This provides the rationale for performing studies of diagnostic accuracy.

## **Chapter 3 Pain Provocation and Pain Intensity During the Application of Physical Tests**

This Chapter relates specifically to Question 1 of this thesis:

*What is the within and between-session intra-examiner reliability of pain provocation and of reports of pain intensity during the application of physical tests to the symptomatic and the asymptomatic hip in people with unilateral hip pain?*

### **3.1 Introduction and background**

Pain felt in the groin, thigh and/or buttock might originate from the hip joint or the soft tissue structures surrounding the hip. Alternatively, it may be referred from the lumbar spine or sacroiliac joint. Tests performed as a part of a clinical examination help identify the likely source of a patient's pain (Laslett, Aprill, McDonald, & Young, 2005; Woolf, 2003). Pain provocation tests (PPT's) are tests applied to load specific structures. A positive test is considered to be one that reproduces the pain that a patient 'normally' feels in association with their presenting condition (i.e. a 'familiar pain'). Reproduction of familiar pain provides the examiner with some evidence that the structure(s) being stressed by that test is the source of the patient's pain (Laslett et al., 2005). A negative test suggests that structure is not the cause of the patient's pain.

Ideally, a PPT should be specific (negative in the absence of pathology) and sensitive (positive in the presence of pathology). However, the application of physical loads to 'normal' structures is provocative if the load is sufficient. Although the requirement for a positive test may be to 'reproduce familiar pain', it would be beneficial to know the prevalence of painful responses reported during the application of PPTs in the asymptomatic population. A low prevalence in this population would allow the clinician to be more confident that a painful response indicates the presence of pathology.

It is not uncommon for clinicians to ask patients to 'score' the intensity of pain reproduced with PPT's. A decrease in pain intensity reported with a test after an intervention compared to that reported prior, suggests that the intervention has had a beneficial effect on the patient (Jensen, Turner, Romano, & Fisher, 1999). This can be seen as an indicator that the correct pathology has been identified (diagnostic) or that an appropriate intervention has been utilised (or both). Similarly, significant changes in pain intensity after a diagnostic anaesthetic injection provides evidence that the structure(s) anaesthetised is the source of a patient's pain.

Thus, high reliability of patient reports of pain reproduction and of pain intensity is important for the correct interpretation of the effect of diagnostic or therapeutic interventions. However, whilst some studies (Laslett & Williams, 1994; Michener, Walsworth, Doukas, & Murphy, 2009; Prather et al., 2010; Wainner et al., 2003) have been undertaken to determine the reliability of pain reproduction during the application of PPTs tests, it appears that the reliability of reports of the intensity of pain provoked by such tests has not been considered.

A secondary focus of this chapter was to consider the prevalence of painful responses to pain provocation tests in both the symptomatic and asymptomatic hips in the cohort of patients recruited for the reliability study. The chapter begins with a literature review that first considers previous research that has examined the reliability of pain reproduction and reports of pain intensity. Next, literature that has reported the prevalence of painful responses during pain provocation tests is presented. A narrative review of research that considers the key methodological factors important to the design and conduct of reliability studies is also provided. Finally, the chapter presents the methods, results and discussion for the reliability study undertaken to answer question 1 of this thesis.

This study is both novel and necessary. Evidence that reports of pain provocation and pain intensity are reliable will allow clinicians and researchers to confidently interpret the findings of such tests.

### **3.2 Literature review**

The general aim of this review was to determine if any previous research in this area had been conducted. Relevant research was used to inform the study reported in this chapter. The specific aims of the review were to answer the following questions:

- What is the *intra*-examiner reliability of reports of pain reproduction and pain intensity resulting from the application of pain provocation tests to the hip?
- What is the *inter*-examiner reliability of reports of pain reproduction resulting from the application of pain provocation tests to the hip?
- What is the prevalence of painful responses resulting from the application of pain provocation tests to the hip?

- What key methodological issues need to be considered in the design and analysis of a study designed to answer the above questions?

The initial search was performed using the search strategy detailed in Chapter 2 in May 2010, prior to the commencement of data collection for the current study, and this was updated with a follow-up search performed in March 2015. Key concepts were identified and searched separately in 5 main categories summarised as: 1) "hip joint" OR "hip pain" OR "groin pain" OR groin OR hip; 2) reliability OR accuracy OR consistency OR validity; 3) pain OR "pain intensity" OR "pain provocation" OR "pain reproduction"; 4) "Physical examination" OR "tests" OR "clinical examination" OR "objective examination" OR "impairment"; 5) prevalence OR incidence OR "cross sectional studies" OR epidemiology.

Results are detailed below under sub-headings that reflect the above aims. No relevant systematic reviews were identified. Hence, only the findings of experimental research are provided.

### **3.2.1 Intra-examiner reliability of reports of pain reproduction and pain intensity**

It appears that there is only one study (Cliborne et al., 2004) that has examined the reliability of reports of pain intensity associated with pain provocation tests for the hip. In this study, four physical tests for the hip (squat, FABER, end ROM hip flexion and 'scour') were applied to patients with symptomatic *knee* osteoarthritis (OA). Patients rated the intensity of pain provoked by the tests using the numeric pain rating scale (NPRS). Tests were repeated, in the same order, after a two-minute rest period. Whilst these authors reported excellent reliability (Intraclass Correlation Coefficient's between 0.87 and 0.90) for reports of pain intensity, this study had some limitations. Firstly, the time between test sessions was only two minutes. It is likely that participant's ratings of pain intensity during the re-testing session would be influenced by their recall of the intensity that they had reported just a few minutes earlier. Secondly, although pain intensity was rated during the application of hip tests, the included participants were patients with symptomatic knee OA. The authors did not make it clear if participants had any hip pain or if a positive test involved provocation of knee or hip pain. This makes it difficult to interpret the results of this study and to generalise them to patients with painful hips.

### 3.2.2 Inter-examiner reliability of reports of pain reproduction

Three studies (Cibere et al., 2008; Martin & Sekiya, 2008; Ratzlaff et al., 2013) that have investigated the reliability of pain provocation during the application of physical tests to people with symptomatic hips were identified. Martin and Sekiya (2008) investigated the within-session, inter-examiner reliability of flexion abduction external rotation (FABER), log roll and flexion adduction internal rotation (FADDIR) tests. This study included 70 subjects with hip pain, recruited from one orthopaedic surgeon. Participants were examined by the surgeon and then by a physiotherapist, blinded to the findings of the orthopaedic surgeon. Examinations were one hour apart. These authors reported kappa values and 95% confidence intervals of 0.63 (0.43-0.83) for FABER, 0.58 (0.29-0.87) for FADDIR and 0.61 (0.41-0.81) for the log roll test. One significant limitation of this study was that a positive test was considered to be one that “reproduced pain in *any* location”. It was not made clear if this meant that the pain had to be similar in site and nature to the pain that the patient normally feels in association with their painful hip (‘a familiar pain’) or if any pain was considered.

Cibere et al. (2008) investigated the reliability hip joint tests and the effect of standardisation of the physical examination in a small group of volunteers (n=6) with symptomatic osteoarthritis of the hip. A positive test was described as “groin pain present”. Six examiners (rheumatologists and orthopaedic surgeons) performed all tests after a two-day standardisation process. These authors used prevalence and bias adjusted kappa (PABAK) to calculate point estimates of reliability. They reported inter-examiner reliability for pain reproduced with end ROM flexion (PABAK = 0.82), external rotation at 90<sup>0</sup> flexion (0.72); internal rotation at 90<sup>0</sup> (0.52); FABER (0.8) and log roll (0.88). With the exception of internal rotation at 90<sup>0</sup>, these results demonstrate reliability values acceptable for clinical use. However, with such a small cohort, the degree of error around these point estimates is likely to be high.

The most recent study (Ratzlaff et al., 2013) investigated the inter-examiner reliability of a number of physical tests for femoroacetabular impingement (FAI) including the log roll, FABER, FADDIR, flexion internal rotation (FIR), full flexion adduction internal rotation (FFADDIR) and flexion adduction compression (FADC). Nine assessors examined both hips of the 12 participants in this study. Five participants were asymptomatic (no history of hip pain), 4 had bilateral hip pain and the remaining 3 had unilateral hip pain (a total of 11 symptomatic hips out of the 24 hips examined). Examinations were 5 minutes apart. A positive test was “pain in the upper thigh, inner

thigh or groin” (not necessarily reproduction of familiar pain). These authors reported positive and negative agreement as well as overall raw agreement (ORA). Negative agreement ranged from 0.68 (0.27, 0.91) for FFADDIR to 1.00 (0.98, 1.0) for log roll. Positive agreement ranged from 0.00 (0.0, 0.25) for log roll to 0.84 (0.71, 0.94) for FFADDIR. Ratzlaff et al. did not use a chance-corrected measure of reliability (e.g. kappa or PABAK). Positive and negative agreement do not correct for chance. However, when both are satisfactorily large, concerns about chance related inflation of reliability are reduced (Cicchetti & Feinstein, 1990). The extremely low positive agreement reported for the log roll test in this study is likely to be a poor indication of the actual reliability of this test. Instead, it reflects the fact that the prevalence of a positive result was just 1%. Ratzlaff and colleagues reported that 6 of the 8 tests examined had ORA of >0.75 and concluded that this indicated adequate inter-examiner reliability and was sufficient to allow clinicians to agree on distinguishing normal hips from those with painful FAI.

One study (Prather et al., 2010) investigated inter-examiner reliability of the straight leg raise, FABER, FADDIR and log roll tests in 28 *asymptomatic* participants (mean age 31 years; SD 11). Multiple examiners (including 9 physiotherapists and 2 orthopaedic surgeons) performed each test once, on both hips, of all participants. A positive test was defined as a report of pain “in the groin, lateral hip or posterior pelvic region”. Prather et al. (2010) were unable to determine reliability of these tests as the prevalence of positive tests in their study was very low, not surprising given that the participants did not have symptomatic hip pathology.

### **3.2.3 Prevalence of painful responses during the application of pain provocation tests**

The aforementioned study by Prather et al. (2010) appears to be the only study that has investigated the prevalence of positive test findings from the application of hip tests to *asymptomatic* participants. In this study, FADDIR was positive in 7.1% of participants and FABER in 3.6%. The log roll test was not provocative in any participant.

In contrast, a number of studies have reported the prevalence of positive test results in people with hip pain (Burnett et al., 2006; Clohisy et al., 2009; Hananouchi, Yasui, Yamamoto, Toritsuka, & Ohzono, 2012; Martin et al., 2008; Maslowski et al., 2010; Mitchell et al., 2003; Narvani, Tsiridis, Kendall, Chaudhuri, & Thomas, 2003; Springer et al., 2009; Suenaga et al., 2002; Troelsen et al., 2009). Of these studies, 6 were

prospective, with 3 performing tests on patients scheduled for arthroscopy (Clohisy et al., 2009; Springer et al., 2009; Suenaga et al., 2002), two on patients scheduled for MRI (Martin et al., 2008; Narvani et al., 2003) and one prior to a guided anaesthetic injection of the hip (Maslowski et al., 2010). Retrospective studies were performed after arthroscopy in three studies (Burnett et al., 2006; Mitchell et al., 2003; Troelsen et al., 2009) and after MRI in one study (Hananouchi et al., 2012). All studies bar one (Suenaga et al., 2002) provided a definition of a positive test. Whilst the majority considered a positive test to be ‘production of groin pain’ (Burnett et al., 2006; Clohisy et al., 2009; Hananouchi et al., 2012; Narvani et al., 2003; Springer et al., 2009; Troelsen et al., 2009), three studies considered that ‘reproduction of groin pain’ to be a positive test (Martin et al., 2008; Maslowski et al., 2010; Mitchell et al., 2003). Across these studies seven physical tests were investigated. Table 3.1 provides detail of the reported prevalence.

**Table 3.1 Prevalence of positive results for physical tests in symptomatic hips (%)**

Study	FFIR	FFER	FABER	FIR	FADDIR	Scour	FADC	Log Roll
Suenaga et al 2002	38	27	-	-	-	-	-	-
Mitchell et al 2003	-	-	-	-	88	-	-	-
Narvani et al 2003	-	-	-	-	-	-	61	-
Burnett et al 2006	-	-	-	-	95	-	-	-
Martin et al 2008	-	-	70	-	83	-	-	-
Clohisy et al 2009	-	-	69	88	-	-	-	30
Springer et al 2009	-	-	-	-	-	89-97	-	-
Troelsen et al 2009	-	-	38	55	-	-	-	-
Maslowski et al 2010	-	-	78	88	-	62	-	-
Hananouchi et al 2012	-	-	-	-	44	-	-	-

FFIR, Full Flexion Internal Rotation; FFER, Full Flexion External Rotation; FABER, Flexion Abduction External Rotation; FIR, Flexion Internal Rotation at 90° Flexion; FADDIR, 90° Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression.

Comparison of prevalence rates across these studies is difficult due to a number of differences between studies including, but not limited to, the manner of performing the tests, the definitions of a positive test, the type and stage of the included pathologies and the background and experience of the examiner. Only four tests were investigated by more than one study. Table 3.1 demonstrates that the prevalence of positive tests results reported for these tests was dissimilar across studies.

### **3.2.4 Methodological considerations**

Various methodological factors that need to be considered in the design and analysis of the reliability study reported in this chapter, were identified in the current search. Key factors and presented below.

#### **Participant and examiner characteristics**

The characteristics of the participants included in a study, and those of the examiners performing the tests, have the potential to influence the results of that study and therefore the generalizability of those results to other settings. Of the 5 relevant reliability studies identified by the current review, all described the method of recruitment, the selection criteria and the included participant characteristics. Similarly, all described the characteristics of the examiners, with four including physiotherapists (Cliborne et al., 2004; Martin & Sekiya, 2008; Prather et al., 2010; Ratzlaff et al., 2013), three including orthopaedic surgeons (Cibere et al., 2008; Martin & Sekiya, 2008; Prather et al., 2010) and two including rheumatologists (Cibere et al., 2008; Ratzlaff et al., 2013).

#### **Measurement of pain intensity**

Whilst a variety of tools can be used to measure pain intensity, some appear to have advantages over others (Hawker, Mian, Kendzerska, & French, 2011). Two of the most commonly used measures, the visual analogue scale (VAS) and the NPRS, have proven reliability and construct validity (Hawker et al., 2011). A limitation of the VAS is that it requires the participant to physically mark a point that represents their pain intensity on a 100mm long continuous line anchored by two verbal descriptors (commonly 'No pain' and 'Worst possible pain'). Good visuospatial abilities are necessary for participants to make this judgement and there is evidence that people with cognitive or motor impairments can have more difficulty with the VAS than the less complicated NPRS (Hawker et al., 2011; Salaffi, Ciapetti, & Carotti, 2012). The 0-10 numbering on the NPRS makes it a pragmatic and easily understood measurement tool that can be administered verbally or graphically (Hawker et al., 2011). The NPRS was used by Cliborne et al. (2004), the only study identified in the current search that has investigated reliability of pain intensity during the application of physical tests.

For the follow-up diagnostic accuracy study, it is important that meaningful changes in pain intensity can be recognised. Several authors have investigated the minimal clinically important difference (MCID) for changes in pain intensity when using the

NPRS (Childs, Piva, & Fritz, 2005; Farrar, Young, LaMoreaux, Werth, & Poole, 2001; Salaffi, Stancati, Silvestri, Ciapetti, & Grassi, 2004). A MCID of 2 points has been demonstrated in patients with sub-acute low back pain (Childs et al., 2005) and chronic pain associated with osteoarthritis (Salaffi et al., 2004). Farrar et al. (Farrar et al., 2001) reported that a 1.74-point reduction from the baseline score was associated with meaningful change (much or very much better) for patients with chronic pain associated with a variety of conditions including diabetic neuropathy, post-herpetic neuralgia, chronic low back pain, fibromyalgia, and osteoarthritis. Farrar and colleagues also reported that the specificity of a 2-point (or 30%) reduction in pain intensity being associated with meaningful change was  $\approx 80\%$  and that specificity increased to  $\approx 92\%$  when there was a 3-point (or 50%) reduction. Salaffi et al. (2004) demonstrated that the intensity of an individual's baseline pain influenced the size of the change necessary to see a MCID. They reported that a meaningful decrease (much better) was 0.7 points (a 17% decrease) for patients with baseline pain of 4 or below. However, a 2.1 point ( $\approx 33\%$ ) decrease was necessary for patients with baseline pain intensity between 4.1 and 7 and this rose to 2.8 points ( $\approx 40\%$ ) for those with baseline intensity greater than 7.

### **Definition of a positive test**

In respect to the current study, a clear definition of a positive test is required if the reliability of pain provocation is to be determined. The application of physical tests to a joint can be provocative even in the absence of any pathology (Prather et al., 2010). Previous researchers in this area have defined a positive test as one that 'reproduces' the patient's pain (Cadogan, McNair, Laslett, & Hing, 2013; Laslett et al., 2005; Sutlive et al., 2008). Some advocate that the term 'reproduction' itself should be clarified, suggesting that only the provocation of a pain that is very similar in nature and site to the pain that it typically experienced with the presenting condition (i.e. a 'familiar' or 'typical' pain) should be considered (Cadogan et al., 2013; Laslett et al., 2005). Various definitions of a positive test were used in the reliability studies identified in the current search including: "pain in the groin, lateral hip or pelvic region" (Prather et al., 2010); "pain in any location" (Martin & Sekiya, 2008); "discomfort in or around the hip" (Cibere et al., 2008); and "pain greater than 1 point on the NPRS" (Cliborne et al., 2004).

With respect to the diagnostic accuracy study that follows, baseline pain needs to be considered so that changes in pain intensity can be interpreted appropriately. The International Spine Intervention Society (ISIL) guidelines suggest that the level of

baseline pain needs to be of sufficient intensity that any change in pain intensity after injection of an anaesthetic is ‘credible and meaningful’ and recommend a minimum of 20mm on a 100mm VAS (Bogduk, 2004b). Whilst these recommendations were made for studies of the spine and sacroiliac joint, they have been applied to diagnostic accuracy studies of peripheral joints including the shoulder and acromio-clavicular joints (Cadogan et al., 2013).

### **Performance of tests**

Whilst all of the studies identified in the current literature search provided detail regarding the performance of the tests that they investigated, it is clear that there is inconsistency in the method of application of many commonly used PPT’s. The performance characteristics of a test will depend not only on the peculiarities of the included participants but also on the manner in which that test is applied and interpreted (Fritz & Wainner, 2001; Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003). If the findings of a study are to be generalized to clinical practice (or compared to other research findings), the method of application needs to be described in sufficient detail that included tests can be replicated (Kottner et al., 2011; Lucas, Macaskill, Irwig, & Bogduk, 2010).

### **Blinding**

Test findings can be influenced by additional information that is not normally part of the test under investigation. Blinding examiners from previous test scores as well as from other background information such as clinical notes, previous imaging/investigations, provisional or confirmed diagnoses, will enhance the independence of the test findings (Lucas et al., 2010). All of the reliability studies identified by the current search reported blinding of examiners from previous test measurements except for (Ratzlaff et al., 2013). Only two studies reported that examiners were blinded from clinical information (Martin & Sekiya, 2008; Ratzlaff et al., 2013).

### **Statistical analysis**

There are various statistical approaches that can be adopted to estimate reliability (Dunn, 1992; Haas, 1991; Hopkins, 2000). There seems to be a general consensus in the literature that continuous data should be analysed with the intraclass correlation statistic (or similar) and categorical data should be analysed with the kappa statistic (or similar) (Cicchetti & Feinstein, 1990; Kottner et al., 2011; Lucas et al., 2010). Similarly, it is

recommended that combinations of coefficients be reported and confidence intervals constructed around point estimates so that readers can develop a better overall understanding of the level of reliability (Kottner et al., 2011; Lucas et al., 2010). Of the studies identified in the current literature search, four used these statistics to estimate reliability. Ratzlaff et al. (2013) reported overall agreement and proportion of specific agreement for categorical data and the median of the absolute difference for continuous variables.

The kappa statistic is a chance corrected measure of reliability. It assumes that some of the agreement that occurs between trials is purely a result of chance and corrects for this possibility (Feinstein & Cicchetti, 1990). Kappa is adversely affected by prevalence (where raters *agree* that the proportion of positive tests and negative tests is different) and bias (the extent to which there is *disagreement* between raters in the proportion of positive and negative results) (Byrt, Bishop, & Carlin, 1993; Fritz & Wainner, 2001; Haas, 1991). In this case it is inappropriate to use kappa and instead prevalence & bias adjusted kappa (PABAK) is justified (Byrt et al., 1993). Therefore, statistical tests should be performed to determine if either prevalence or bias is likely to affect the calculated kappa.

Parameters like kappa and ICC are relative measures of reliability and relate variability in participants to error associated with measurement. They reflect the degree to which individuals maintain their 'position' or 'ranking' in regard to the repeated measurements (Atkinson & Nevill, 1998). Measures of absolute reliability include the standard error of measurement (SEM), limits of agreement (LOA) and smallest detectable difference (SDD). These measures estimate the size of error involved with repeated measurements (in the same metric as the original scale) and hence allow the clinician to consider how important that error might be when interpreting change in scores for an individual patient. Test results smaller than the SEM are unlikely to be detected reliably in clinical practice (Knols, Aufdemkampe, De Bruin, Uebelhart, & Aaronson, 2009).

### **3.2.5 Review Summary**

Whilst both intra-examiner and inter-examiner reliability are important, intra-examiner reliability is essential to this research. The proposed diagnostic accuracy study (Chapter 5) requires the researcher to perform tests on participants before and soon after the administration of an intra-articular injection of anaesthetic into the participant's painful hip joint. Response to the injection will be determined by the percent reduction in pain

produced with these tests. Thus, the researcher needs to have confidence in the reliability of reports of pain reproduction and of pain intensity during the application of such tests applied by the researcher.

It is clear that there is insufficient research evidence to demonstrate that these measures are reliable. The existing literature has not considered the reliability of ratings of pain intensity. Whilst the reliability of pain reproduction has been reported, this research has primarily investigated inter-examiner reliability. This, along with the considerable heterogeneity of this research and the limitations presented above make it inappropriate to assume that such tests will have sufficient reliability to justify their inclusion in main study of this thesis.

Therefore, the main purpose of the current study was to determine the within and between-session reliability of pain reproduction and reports of pain intensity during the application of pain provocation tests for the hip in a group of patients with hip pathology. A secondary aim was to determine the prevalence of positive tests in both symptomatic and asymptomatic hips. The results of this study were used to determine which tests and measures were included in the diagnostic accuracy study. The reliability study also served as a 'pilot' for trialling methods of recruitment and data collection that were to be used in the diagnostic accuracy study.

### **3.3 Methods and procedures of the current study**

#### **3.3.1 Study design**

This study was a test-retest study of the between and within-session, intra-examiner reliability of pain provocation tests for the hip joint.

#### **3.3.2 Sample size calculation**

A minimum sample size of 18 participants was calculated, based on the aim of detecting a minimum acceptable reliability of .60 with an  $\alpha$  level (1-tailed) of .05 and power of 80% (Sim & Wright, 2005).

#### **3.3.3 Participants**

Patients with pain felt primarily in the groin or deep buttock region, which had been present for a minimum of one month, were recruited from selected sports medicine practices as well as by advertisements placed in a University setting. Prospective participants were screened to ensure that the site and nature of their pain suggested hip

pathology and not pathology that might mimic hip pain e.g. pain referred from the lumbar spine or sacroiliac joint. Inclusion criteria were: a history of at least one month of pain felt primarily in the groin or deep buttock region, between 20 and 80 years old and able to speak English. Exclusion criteria were: a history of hip joint replacement, treatment for low back pain within the preceding 12 months, pregnancy, illness or systemic disease. Also, prospective participants who described their pain as severe (or greater than 9 on the NPRS) were excluded as a requirement for ethical approval. Ethical approval for this study was granted by the AUT University Committee (AUTEK) (Reference number 10/44) (See Appendix 1).

### **3.3.4 Examiner**

One experienced examiner (the PhD Candidate), a physiotherapist with 32 years experience, assessed all participants.

### **3.3.5 Procedures**

Potential participants were provided with a study information sheet that explained the purpose of the research, eligibility criteria, experimental procedures, possible associated discomfort or risks, compensation and privacy issues and procedures for dealing with any concerns about the study (see Appendix 2). They were given time to read and consider this information sheet and to ask for advice from third parties if necessary. The researcher followed up a day or so later to determine if they were interested in participating in the study and to answer any questions they had about the study. Those that expressed a desire to be included in the study were then screened to ensure that they met all inclusion criteria and none of the exclusion criteria. Appointments for data collection were made for those that were appropriate to include in the study.

On the day of data collection, participants signed a statement of informed consent (see Appendix 3). Baseline data including age, occupation, levels of activity, cause and nature of current symptoms, associated symptoms, aggravating and easing factors and any previous injury details were collected via a standardised questionnaire (see Appendix 4). This questionnaire was developed by the researcher on the basis of current understandings in the field (Clohisy et al., 2009; Cook, 2010; Domb et al., 2009; Martin et al., 2010b; Peterson et al., 1992; Woolf, 2003) and on expert opinion. This opinion was sort by interviewing five medical specialists with expertise in the diagnosis and management of patients with hip joint pain (two orthopaedic surgeons, two sports physicians and a recognised specialist physiotherapist). Also, for the sake of

completeness, four patients with hip pain were interviewed to get a patient perspective of the diagnostic process. These interviews helped to identify items from the history and clinical examination that experts and patients considered important. Consent for these interviews were approved as a part of the current reliability study (see Appendix 5 and Appendix 6 for detail). A body chart was used for participants to indicate the site of pain (both primary and secondary) and any associated symptoms. Participants also completed validated questionnaires that provided information about activity levels (the Lower Limb Task Questionnaire) (McNair et al., 2007) (see Appendix 14), and the possibility of a neuropathic pain component to the participant's problem (The self-report version of the Leeds Assessment of Neuropathic Symptoms and Signs questionnaire [S-LANSS]) (Bennett, Smith, Torrance, & Potter, 2005; Weingarten et al., 2007) (see Appendix 15). Finally, a composite baseline pain score was calculated for each participant by averaging the pain intensity scores that they reported for pain felt in the morning, afternoon, evening and with activity over the 2 days immediately prior to data collection.

Standardised versions of physical tests were performed bilaterally on each participant (for detail regarding performance of each test see Appendix 7). To minimise order effects, tests were performed in a random order (pre-determined by a random number generator) for each participant. For participants with unilateral hip pain, the asymptomatic hip was tested first. For those with bilateral hip joint pain, the least painful hip was tested first. Participants were required to report the onset of any pain during the application of the tests. When the symptomatic hip was being tested, participants were asked if the test provoked a 'familiar' pain i.e. pain that was very similar in nature and site to the pain they usually felt with their hip. For the non-symptomatic hip, the patient was asked to describe where the pain was felt. The participant was then asked to rate the intensity of the pain provoked using the 11-point NPRS. This scale was anchored at 0 (no pain) and 10 (worst pain imaginable). Results for each test were recorded immediately on a standardised assessment form.

Three trials were performed. Trial one was followed approximately 60 minutes later by trial two. Trial three was 2-7 days later (dependent on participant availability). During the interval between trial two and three, participants were asked not to make any changes to their normal daily routine in terms of mechanical loading on the hip (exercise, work activity) or to vary the dose of any medications that they had been taking for the two days prior to the first day of testing. When participants returned for

trial three, they completed a questionnaire that explored these factors so that anything that might explain differences in pain intensity from baseline testing could be identified (see Appendix 8). Also, the composite baseline pain score calculated prior to trial one was repeated on the day of testing for trial three so that any changes in baseline pain intensity could be identified.

### **3.3.6 Blinding**

The examiner was blinded to all baseline data until after the physical examination except for detail regarding the exact site, the nature and intensity of pain. Knowledge of these details was necessary so that any pain produced by the test procedures could be recognised and therefore confirmed as the same pain as the patient typically feels. On completion of each trial, results for that trial were placed in a coded, sealed envelope. To minimise bias, neither the examiner nor the participant could refer back to the previous results.

### **3.3.7 Included tests**

There are a number of studies that have reported the diagnostic accuracy of various tests for the hip (Burgess et al., 2011; Kivlan, Martin, & Sekiya, 2011; Martin et al., 2008; Maslowski et al., 2010). Tests that have been commonly investigated include: the impingement test, the scour/quadrant test and FABER. Whilst these tests are widely used in clinical practice, an examination of the relevant literature reveals that there is inconsistency in the description and application of these tests. To provide a comprehensive report on the reliability of these tests, we investigated a number of the modified versions of tests as reported in the literature and as described by various expert clinicians. Appendix 7 provides full descriptions of the tests included in this study.

### **3.3.8 Definition of a positive test**

A 'positive' test for the painful hip was considered to be reproduction of 'familiar' pain, provided the pain intensity was greater than or equal to 2 points on the NPRS. A positive test for the asymptomatic hip was considered to be pain greater than 2 points on the NPRS felt anywhere in the hip region (groin, greater trochanteric or deep buttock regions).

### **3.3.9 Analysis**

Intra-examiner reliability for categorical data (pain produced or not) was calculated using Cohen's chance-corrected Kappa. Cohen's Kappa values were calculated for each

test on the basis of whether or not a ‘familiar’ pain was reproduced i.e. the reproduction of pain was considered and not pain intensity. Statistical tests were employed to determine if either prevalence or bias had affected the calculated kappa. For these tests, a zero value for either the Prevalence Index (PI) or Bias Index (BI) was interpreted as indicating that no prevalence or bias was present (Byrt et al., 1993). A maximum BI value of 1 was considered as indicating significant bias. For PI values (which range from -1 to +1), scores of greater than 0.5 or less than -0.5 were considered to indicate significant prevalence issues existed. These values were used to determine when to use the PABAK statistic (Cadogan, Laslett, Hing, McNair, & Williams, 2010; Sim & Wright, 2005).

Pain intensity data were examined to determine if they exhibited normal distribution using the Kolmogorov-Smirnov non-parametric test. Variables with a significance of greater than 0.05 were classified as having a normal distribution and were then assessed for reproducibility using single-measure Intraclass Correlation Coefficient's (ICC<sub>2,1</sub>) (two way random and absolute) via the Statistical Package for the Social Sciences software, version 19.0 (SPSS Inc, Chicago, USA). The standard error of measurement (SEM) for each test was calculated so that the magnitude of any error associated with estimations of pain intensity could be appreciated. We used the formula  $SEM = SD \times \sqrt{1-R}$  where the standard deviation (SD) was obtained from the data from Trial 1 and the reliability (R) was the reliability value as calculated in the current study for each test (Wyrwich, 2004). Variables that did not exhibit normal distribution were assessed for reproducibility using Lin's Concordance Correlation Coefficient (CCC) (Dunn, 1992; Lin, 1989) via the online calculator provided by the National Institute of Water and Atmospheric Research (<http://www.niwa.co.nz/online-services/statistical-calculators/concordance>).

Ninety five per cent confidence intervals (CI) were constructed as a measure of precision for both the kappa and ICC/CCC values (Sim & Wright, 2005). Reliability values for ICC's, Lin's CCC, Kappa and PABAK were interpreted according to the guidelines of Landis and Koch i.e. <0.00 = poor; 0.00 - 0.20 = slight; 0.21 - 0.40 = fair; 0.41 - 0.60 = moderate; 0.61 - 0.80 = substantial; 0.81 - 1.00 = almost perfect (Landis & Koch, 1977).

### 3.4 Results

Eighteen volunteers (11 females) with painful hips were recruited; mean age 29.5 years (range 20-51); mean body mass index 24.9 BMI kg/m<sup>2</sup> (range 17.9-35.9). All participants completed all three trials. The average number of days between trial two and three was four. Two participants had bilateral hip joint pain. For these subjects, the 'most' painful hip was considered as the 'symptomatic' hip. The data from their 'less' painful hip was recorded but not included in the analysis. As these hips were painful, it would have been inappropriate to include these data with that from asymptomatic hips. It would also be inappropriate to classify these hips as a second 'symptomatic' hip as this could violate the assumption of independence of observation required in statistical analysis. At the time of assessment, one participant had a constant low intensity pain (NPRS of 2). For that participant a positive test was considered to be one that caused an increase from her baseline pain. For her, the NPRS was anchored at 2 (baseline pain) and 10 (worst pain imaginable).

Most participants (78%) were employed at the time of data collection. Diagnoses for the participants included labral tear (n = 7), osteoarthritis (n = 6), femoro-acetabular impingement (n = 4), inflammatory arthritis (n = 2) and ligamentum teres rupture (n = 1). Fourteen diagnoses were confirmed with appropriate medical imaging (radiograph, magnetic resonance imaging CT) and/or blood tests. The referring physician made a diagnosis on the basis of clinical findings for four participants.

The average duration of symptoms at the time of data collection was 25.2 months with a range between 1 and 60 months. Table 3.2 provides detail regarding the frequency and nature of participants' symptoms. Aching pain was the main symptom for the majority (83%) of participants. The most common associated symptom was crepitus (78%).

**Table 3.2 Frequency of reported symptoms**

<b>Symptom</b>	<b>Number of Participants</b>	<b>%</b>
Main Symptom		
Pain	15	83
Stiffness	3	17
Main Pain Nature		
Sharp	3	17
Ache	15	83
Associated Symptoms		
Crepitus	14	78
Painful click	7	39
Non-painful click	9	50
Locking	5	28
Tingling	4	22
Giving way	4	22
Burning or coldness	3	17
Pins & needles	3	17
Loss of movement	10	55
Morning stiffness	7	39

All but two subjects felt pain in their groin and 78% of participants reported that the groin was the site of their predominant (or main) pain. Forty percent of participants had pain in the region of the greater trochanter and 45% had buttock pain. Associated pain was also reported in the anterior and/or posterior aspects of the thigh (see Table 3.3). When a test reproduced pain, that pain was felt at the same body site at subsequent trials for all participants. Pain was reported at the primary site for all participants and all tests except for three participants who also experienced pain at a secondary site for a small number of tests. Five participants reported that a test caused pain that they could not confidently say was a ‘familiar pain’. This occurred with only one or two tests for three participants and with four tests for the others.

**Table 3.3 Region of pain**

<b>Region</b>	<b>Number of Participants</b>	<b>%</b>
Groin	16	89
Upper thigh	2	11
Anterior thigh	1	5
Knee	1	5
Mid-buttock	5	28
Low-buttock	3	17
Upper hamstring	3	17
Posterior thigh	1	5
Trochanteric Region	7	40

Table 3.4 provides detail regarding the behaviour of pain. All participants reported having pain during the month prior to participation in the study and just over one half

(55%) of the participants had experienced pain on most days during that month. All but two participants reported having pain free days during that same period. The range of ‘worst’ pain intensity during the month was from 1 to 8 with a mean of 5.3 (SD 1.9). One participant described constant pain. Three participants (17%) regularly used medication for pain relief.

**Table 3.4 Pain behaviour**

	Number	%
Over the last month		
Pain on most days	10	55
Painful to sleep on sore side	3	17
Painful to sleep on good side	2	11
Wakes during the night with pain	6	33
Intermittent Pain	17	94

The most common aggravating activity was walking or using stairs, with 72% of participants reporting pain with these activities (see Table 3.5 for detail). Jogging and twisting were the most provocative activities.

**Table 3.5 Pain intensity (NPRS)**

Aggravating Factor	Number of Participants (%)	Maximum Intensity	Mean Intensity	Std. Deviation
Time of Day				
Morning <sup>1</sup>	10 (55)	6	2.8	1.6
Afternoon <sup>1</sup>	13 (72)	7	2.8	1.6
Evening <sup>1</sup>	14 (78)	5	2.9	1.2
Activity				
Walking	13 (72)	5	2.6	1.7
Stairs	13 (72)	3	1.8	0.9
Standing	9 (50)	6	2.7	1.7
Sit to Stand	8 (44)	6	3.0	1.7
Getting in/out of car	9 (50)	7	3.1	2.0
Putting on socks	8 (44)	7	3.2	2.4
Squatting	12 (67)	5	2.2	1.5
Driving	8 (44)	5	2.8	1.7
Jogging/running	12 (67)	8	3.6	2.3
Twisting	12 (67)	9	3.8	2.5

<sup>1</sup> Mean pain intensity felt over the 2 days prior to participation

The majority of participants were unsure about the cause of their hip pain with only 33% attributing the cause to a specific incident. Eighty-nine percent played regular sport however ten (55%) had to limit their participation in sport because of their hip pain. Sixty-seven percent of participants played sport that could be considered ‘high demand’

for the hip joint including running, racquet sports, contact sports, weight lifting and martial arts (see Table 3.6).

**Table 3.6 Cause, history and activity levels**

	Number of Participants	%
Cause		
Trauma	6	33
Overuse	3	17
Unknown	8	44
Previous problem same hip	7	40
Previous problem other hip	3	17
Family history of hip problems	8 <sup>1</sup>	44
Plays sport regularly	16	89
Plays high demand sport	12	67
Sport limited because of hip	10	55

<sup>1</sup> Two participants unsure

Scores on the Self-completed Leeds Assessment of Neuropathic Symptoms and Signs (S-LANSS) questionnaire were generally very low with a mean of just 3.5 out of a possible maximum of 24 (see Table 3.7). One participant scored a maximum of 24 on this questionnaire suggesting that she had a neuropathic component to her pain.

**Table 3.7 Baseline composite pain intensity, functional and neuropathic pain scores**

	Minimum	Maximum	Mean	SD
Composite Pain Intensity Baseline Scores <sup>1</sup>				
Trial 1	0	5.3	1.9	1.4
Trial 2	0	5.3	2.1	1.7
Functional Status (LLTQ)				
ADL score <sup>2</sup>	29	40	37.1	3.8
Recreational score <sup>2</sup>	19	40	32.2	7.9
Neuropathic Pain Status				
S-LANSS score <sup>3</sup>	0	24	3.5	5.5

LLTQ, Lower Limb Task Questionnaire; SD, Standard deviation

S-LANSS, Self-completed Leeds Assessment of Neuropathic Symptoms and Signs;

<sup>1</sup> Group mean score calculated by averaging individual scores for pain felt in morning, afternoon, evening and with activity over the preceding 48 hours; <sup>2</sup> Maximum score =40; <sup>3</sup> Maximum score = 24

There was no statistically significant difference ( $p = 0.09$ ) between the group mean composite baseline scores calculated at trial one and trial three. Only two individuals had a change in this score of greater than 1 point on the NPRS. Both of these individuals had a *decrease* in the intensity of pain experienced over the two days prior to trial 3 compared to the two days prior to trial 1 (data not shown).

### 3.4.1 Reliability of categorical yes/no response to pain reproduction

Table 3.8 provides detail regarding the within-session reliability of tests when a positive test was considered to be reproduction of familiar pain (with an intensity of 2 or greater on the NPRS). One test (passive extension in prone) demonstrated perfect reliability with a PABAK value of 1. Four tests demonstrated ‘almost perfect’ reliability, ten tests demonstrated ‘substantial’ reliability, eleven demonstrated ‘moderate’ reliability and three tests, ‘fair’ reliability. Kappa values range from 0.29 for full flexion external rotation (FFER) to 0.85 for the bent knee fall out (BKFO) test. High prevalence index values (either less than -0.5 or greater than 0.5) were seen with 17 tests. Nine of these tests also had high chance agreement (greater than 0.85). Consequently, PABAK was used as an index of reliability for these tests rather than kappa (Byrt et al., 1993). PABAK values ranged from 0.44 to 1. Percent agreement ranged from 71% for full flexion and to 100% for rise from chair. Seventeen tests had percent agreement scores of 80% or better.

Table 3.9 provides detail regarding the between-session reliability of tests when a positive test was reproduction of familiar pain (with an intensity of 2 or greater). Six tests demonstrated ‘almost perfect’ reliability, fifteen tests demonstrated ‘substantial’ reliability; five demonstrated ‘moderate’ reliability; one test ‘fair’ and one test ‘poor’ reliability.

Kappa values range from 0.03 for adduction in standing (ADDSt) to 0.78 (FABER, FIR, FER). High prevalence index values (either less than -0.5 or greater than 0.5) were seen with 17 tests. Eleven tests also had high chance agreement (greater than 0.85). Consequently, PABAK was used as an index of reliability for these tests rather than kappa (Byrt et al., 1993). PABAK values ranged from 0.44 to 0.89. Percent agreement ranged from 56% for adduction in standing and to 100% for rise from chair and resisted extension. Nineteen tests had percent agreement scores of 80% or better.

Table 3.8 Within-session reliability for yes/no reports of pain reproduction

Test	Kappa (95% CI)	PABAK	Percent (Observed) Agreement
<b>Impingement Tests</b>			
Quadrant		0.78 <sup>1</sup>	0.89 (0.65, 0.99)
FADDIR	0.51 (0.10, 0.91)		0.78 (0.52, 0.94)
FADC	0.56 (0.12, 1.00)		0.83 (0.59, 0.96)
FIRC	0.56 (0.18, 0.93)		0.78 (0.52, 0.94)
<b>Miscellaneous Tests</b>			
FF	0.38 (-0.1, 0.83)		0.71 (0.44, 0.90)
FABER	0.43 (0.01, 0.85)		0.72 (0.47, 0.90)
BKFO	0.85 (0.57, 1.10)		0.94 (0.73, 1.00)
EPr		1.0 <sup>1,2</sup>	1.0
Log Roll		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
<b>Internal Rotation Tests</b>			
FFIR		0.56 <sup>1</sup>	0.78 (0.52, 0.94)
FIR		0.44 <sup>1</sup>	0.72 (0.47, 0.90)
IRSit	0.40 (-0.04, 0.84)		0.72 (0.47, 0.90)
IRPr	0.66 (0.31, 1.00)		0.83 (0.59, 0.96)
IRSt	0.56 (0.18, 0.93)		0.78 (0.52, 0.94)
<b>External Rotation Tests</b>			
FFER	0.29 (-0.16, 0.74)		0.72 (0.47, 0.90)
FER		0.44 <sup>1</sup>	0.72 (0.47, 0.90)
ERSit		0.67 <sup>1,2</sup>	0.83 (0.59, 0.96)
ERPr		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
ERSt		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
<b>'Functional' tests</b>			
ADDSt	0.54 (0.15, 0.93)		0.78 (0.52, 0.94)
ADDKn	0.78 (0.49, 1.10)		0.94 (0.73, 1.00)
FKn	0.73 (0.38, 1.10)		0.89 (0.65, 0.99)
Squat		0.88 <sup>1,2</sup>	0.94 (0.00, 0.00)
Rise		-	1.0
<b>Resisted Tests</b>			
RAD		0.67 <sup>1,2</sup>	0.83 (0.59, 0.96)
RAB		0.67 <sup>1</sup>	0.83 (0.59, 0.96)
RF		0.78 <sup>1</sup>	0.89 (0.65, 0.99)
RE		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
RIR		0.56 <sup>1</sup>	0.78 (0.52, 0.94)
RER		0.56 <sup>1</sup>	0.78 (0.52, 0.94)

CI = confidence intervals; PABAK= prevalence and bias adjusted kappa

<sup>1</sup> Tests with high prevalence index values (either less than -0.5 or greater than 0.5)

<sup>2</sup> Tests with high chance agreement (greater than 0.85)

- Incalculable (all tests were true negatives)

FADDIR, 90° Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FIRC, Flexion Internal Rotation Compression; FF, Full Flexion; FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; EPr, Extension in Prone; FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRPr, Internal Rotation in Prone; IRSt, Internal Rotation in Standing; FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERPr, External Rotation in Prone; ERSt, External Rotation in Standing; ADDSt, Adduction in Standing; ADDKn, Adduction in 4-point kneel; FKKn, Hip flexion in 4-point kneel; Squat, squat to chair; Rise, rise from chair; RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°

**Table 3.9 Between-session reliability for yes/no reports of pain reproduction**

<b>Test</b>	<b>Kappa (95% CI)</b>	<b>PABAK</b>	<b>Percent (Observed) Agreement</b>
<b>Impingement Tests</b>			
Quadrant		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
FADDIR	0.61 (0.21, 1.00)		0.83 (0.59, 0.96)
FADC	0.56 (0.21, 0.90)		0.78 (0.52, 0.94)
FIRC	0.56 (0.18, 0.93)		0.78 (0.52, 0.94)
<b>Miscellaneous Tests</b>			
FF	0.38 (-0.07, 0.83)		0.71 (0.44, 0.90)
FABER	0.78 (0.49, 1.06)		0.89 (0.65, 0.99)
BKFO	0.45 (-0.01, 0.91)		0.78 (0.52, 0.94)
EPr		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
Log Roll		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
<b>Internal Rotation Tests</b>			
FFIR	0.73 (0.38, 1.07)		0.89 (0.65, 0.99)
FIR	0.78 (0.49, 1.07)		0.89 (0.65, 0.99)
IRSit	0.60 (0.19, 1.00)		0.82 (0.57, 0.96)
IRPr		0.67 <sup>1</sup>	0.83 (0.59, 0.96)
IRSt	0.75 (0.44 - 1.07)		0.89 (0.65, 0.99)
<b>External Rotation Tests</b>			
FFER		0.67 <sup>1</sup>	0.83 (0.59, 0.96)
FER	0.78 (0.49, 1.07)		0.89 (0.65, 0.99)
ERSit		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
ERPr		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
ERSt		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)
<b>'Functional' tests</b>			
ADDSt	0.03 (-0.29, 0.35)		0.56 (0.31, 0.78)
ADDKn		0.44 <sup>1</sup>	0.72 (0.47, 0.90)
FKn		0.67 <sup>1</sup>	0.83 (0.59, 0.96)
Squat		0.88 <sup>1,2</sup>	0.94 (0.73, 1.00)
Rise		-	1.0
<b>Resisted Tests</b>			
RAD		0.67 <sup>1</sup>	0.83 (0.59, 0.96)
RAB		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
RF		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
RE		-	1.0
RIR		0.89 <sup>1,2</sup>	0.94 (0.73, 1.00)
RER		0.78 <sup>1,2</sup>	0.89 (0.65, 0.99)

CI = confidence intervals; PABAK= prevalence and bias adjusted kappa

<sup>1</sup> Tests with high prevalence index values (either less than -0.5 or greater than 0.5)

<sup>2</sup> Tests with high chance agreement (greater than 0.85)

- Incalculable (all tests were true negatives)

FADDIR, 90° Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FIRC, Flexion Internal Rotation Compression; FF, Full Flexion; FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; EPr, Extension in Prone; FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRPr, Internal Rotation in Prone; IRSt Internal Rotation in Standing; FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERPr, External Rotation in Prone; ERSt External Rotation in Standing; ADDSt, Adduction in Standing; ADDKn Adduction in 4-point kneel; FKKn, Hip flexion in 4-point kneel; Squat, squat to chair; Rise, rise from chair; RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°

### 3.4.2 Reliability of reports of pain intensity

The results for within-session reliability of ratings of pain intensity are presented in Table 3.10. Six tests demonstrated normal distribution and were assessed for reliability with ICC's whereas the remaining tests were assessed using Lin's CCC. ICC values ranged from 0.64 for FFIR to 0.93 for the quadrant test. CCC values ranged from -0.05 for the log roll test to 0.90 for passive extension in prone (EPr).

Using the Landis and Koch interpretations of reliability scores, seven tests demonstrated 'almost perfect' reliability. Another seven tests scored in the 'substantial' range. Six tests had moderate reliability and seven tests demonstrated fair (or worse) reliability. Reliability could not be calculated for two tests (squat and rise) due to the high number of zero pain intensity scores reported for these tests. Only one person reported pain with either of these tests. Table 3.10 also provides detail regarding the mean intensity and range of pain (using the NPRS) reported for each test across trials one and two (1 hour apart). Mean pain intensity was highest for the quadrant test (3.8). The widest range of scores was 0 to 9, seen with FADC and the narrowest was 0-1 (rise). SEM scores ranged from 0.2 to 1.3 points on the NPRS.

The results for between-session reliability are presented in Table 3.11. Five tests had a normal distribution and were assessed for reliability with ICC's. The remaining tests were assessed using Lin's CCC. ICC values ranged from 0.74 for FADC to 0.85 for quadrant and FIR. CCC values ranged from -0.02 for external rotation in standing (ERSt) to 0.84 for internal rotation in standing (IRSt). The tests with the highest reliability ('almost perfect') were quadrant (ICC=0.85), FIR (0.85) and IRSt (0.84). Seventeen tests exhibited concordance values in the 'almost perfect' or 'substantial' range. Table 3.9 also provides detail in regard to the mean intensity and range of pain (using the NPRS) reported for each test across trials two and three (1 to 6 days apart). Mean pain intensity was highest for the quadrant test (3.8). The widest range of scores was 0 to 9 (for FADC) and the narrowest was 0-2 (resisted abduction). SEM scores ranged from 0.3 to 1.8 with the average SEM across all tests being 0.8 points on the NPRS.

**Table 3.10 Within-session reliability for reports of pain intensity**

Test	CCC (95% CI)	ICC <sub>2,1</sub> (95% CI)	Mean Intensity Trial 1/Trial 2 (range)	SEM
<b>Impingement Tests</b>				
Quadrant		0.93 (0.83, 0.98)	3.8/3.8 (0-8.5)	0.6
FADDIR		0.80 (0.54, 0.92)	2.8/2.9 (0-8)	1.1
FADC		0.86 (0.66, 0.94)	3.1/3.3 (0-9)	0.9
FIRC	0.45 (0.02, 0.75)		1.4/1.9 (0-6)	1.3
<b>Miscellaneous Tests</b>				
FF	0.84 (0.61, 0.94)		2.5/2.5 (0-6)	0.8
FABER	0.57 (0.16, 0.81)		2.3/2.7 (0-8.5)	1.5
BKFO	0.81 (0.62, 0.91)		1.1/1.0 (0-7)	0.9
EPr	0.90 (0.78, 0.95)		0.3/0.3 (0-3.5)	0.2
Log Roll	-0.05 (-0.05, 0.35)		0.1/0.2 (0-3.5)	0.5
<b>Internal Rotation Tests</b>				
FFIR		0.64 (0.26, 0.85)	3.5/3.3 (0-8)	1.0
FIR		0.82 (0.57, 0.93)	2.2/2.2 (0-6)	0.8
IRSit	0.68 (0.33, 0.87)		1.3/1.1 (0-4)	0.8
IRPr		0.82 (0.60, 0.93)	2.5/2.2 (0-8)	1.0
IRSt	0.79 (0.52, 0.91)		1.6/1.4 (0-5)	0.8
<b>External Rotation Tests</b>				
FFER	0.48 (0.08, 0.75)		1.2/0.7 (0-4)	1.2
FER	0.51 (0.12, 0.77)		1.1/0.6 (0-6)	1.3
ERSit	-0.05 (-0.48, 0.40)		0.3/0.3 (0-3)	0.6
ERPr	0.59 (0.20, 0.82)		0.3/0.2 (0-3)	0.4
ERSt	0.33 (-0.11, 0.67)		0.3/0.3 (0-3)	0.7
<b>'Functional' tests</b>				
ADDSt	0.56 (0.14, 0.81)		1.3/1.4 (0-4.5)	0.8
ADDKn	0.75 (0.46, 0.89)		1.8/2.4 (0-6.5)	1.0
FKn	0.75 (0.46, 0.90)		1.2/1.1 (0-7)	0.9
Squat	-		0.1/0.0 (0-2)	-
Rise	-		0.1/0.0 (0-1)	-
<b>Resisted Tests</b>				
RAD	0.18 (-0.27, 0.56)		0.3/0.2 (0-2.5)	0.7
RAB	0.18 (-0.14, 0.46)		0.6/0.1 (0-2)	0.7
RF	0.69 (0.44, 0.84)		0.6/0.3 (0-4)	0.7
RE	0.16 (-0.23, 0.50)		0.1/0.0 (0-1.5)	0.7
RIR	0.17 (-0.26, 0.54)		0.8/0.4 (0-3.5)	1.0
RER	0.16 (-0.23, 0.48)		0.4/0.1 (0-2)	0.7

CCC, Concordance Correlation Coefficient (Lins); ICC, Intraclass Correlation Coefficient; CI, Confidence Intervals; SEM, Standard Error of Measurement; - Incalculable due to zero values

FADDIR, 90° Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FIRC, Flexion Internal Rotation Compression; FF, Full Flexion; FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; EPr, Extension in Prone; FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRPr, Internal Rotation in Prone; IRSt, Internal Rotation in Standing; FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERPr, External Rotation in Prone; ERSt, External Rotation in Standing; ADDSt, Adduction in Standing; ADDKn, Adduction in 4-point kneel; FKKn, Hip flexion in 4-point kneel; Squat, squat to chair; Rise, rise from chair; RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°

**Table 3.11 Between-session reliability for reports of pain intensity**

<b>Test</b>	<b>CCC (95% CI)</b>	<b>ICC<sub>2,1</sub> (95% CI)</b>	<b>Mean Intensity Trial 1/Trial 2 (range)</b>	<b>SEM</b>
<b>Impingement Tests</b>				
Quadrant		0.85 (0.63, 0.94)	3.8/3.1 (0-8)	0.9
FADDIR		0.82 (0.60, 0.93)	2.9/2.5 (0-8)	1.0
FADC		0.74 (0.40, 0.89)	3.3/2.4 (0-9)	1.3
FIRC	0.79 (0.54, 0.91)		1.9/1.7 (0-7)	0.8
<b>Miscellaneous Tests</b>				
FF	0.82 (0.58, 0.93)		2.5/2.6 (0-7)	0.9
FABER	0.65 (0.31, 0.84)		2.7/1.9 (0-8.5)	1.5
BKFO	0.73 (0.43, 0.88)		1.0/1.1 (0-7)	0.8
EPr	-		0.3/0.0 (0-3.5)	NA
Log Roll	-0.05 (-0.46, 0.38)		0.2/0.2 (0-3.5)	0.9
<b>Internal Rotation Tests</b>				
FFIR		0.80 (0.46, 0.93)	3.3/2.5 (0-8)	0.9
FIR		0.85 (0.58, 0.95)	2.2/1.7 (0-6)	0.8
IRSit	0.64 (0.26, 0.84)		1.1/1.1 (0-4)	0.8
IRPr	0.81 (0.57, 0.93)		2.2/1.9 (0-7)	0.9
IRSt	0.84 (0.63, 0.94)		1.4/1.2 (0-5)	0.6
<b>External Rotation Tests</b>				
FFER	0.67 (0.32, 0.86)		0.7/0.6 (0-4)	0.7
FER	0.68 (0.33, 0.86)		0.6/0.8 (0-4.5)	0.8
ERSit	0.42 (0.02, 0.71)		0.3/0.2 (0-3)	0.6
ERPr	0.46 (0.02, 0.75)		0.2/0.3 (0-3)	0.5
ERSt	-0.02 (-0.43, 0.40)		0.3/0.4 (0-3.5)	0.6
<b>'Functional' tests</b>				
ADDSt	0.17 (-0.16, 0.46)		1.4/0.6 (0-4.5)	1.2
ADDKn	0.39 (-0.05, 0.71)		2.4/1.9 (0-6.5)	1.8
FKn	0.73 (0.43, 0.88)		1.1/0.9 (0-7)	1.0
Squat	-		0.0/0.1 (0-2)	NA
Rise	-		0/0 (0)	NA
<b>Resisted Tests</b>				
RAD	0.01 (-0.33, 0.36)		0.2/0.4 (0-3.5)	0.5
RAB	0.43 (0.02, 0.73)		0.1/0.2 (0-2)	0.3
RF	0.63 (0.36, 0.80)		0.3/0.4 (0-5)	0.5
RE	0.14 (0.22, 0.48)		0/0 (0)	0.5
RIR	0.72 (0.40, 0.88)		0.4/0.3 (0-4)	0.4
RER	0.15 (-0.23, 0.48)		0.1/0.3 (0-4)	0.4

CCC, Concordance Correlation Coefficient (Lins); ICC, Intraclass Correlation Coefficient; CI, Confidence Intervals; SEM, Standard Error of Measurement; - Incalculable due to zero values

FADDIR, 90° Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FIRC, Flexion Internal Rotation Compression; FF, Full Flexion; FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; EPr, Extension in Prone; FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRPr, Internal Rotation in Prone; IRSt Internal Rotation in Standing; FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERPr, External Rotation in Prone; ERSt External Rotation in Standing; ADDSt, Adduction in Standing; ADDKn Adduction in 4-point kneel; FKKn, Hip flexion in 4-point kneel; Squat, squat to chair; Rise, rise from chair; RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°

### 3.4.3 Prevalence of positive responses

Table 3.12 provides data regarding the prevalence of positive test responses for each test in each trial for the painful hip as well as the prevalence averaged over the three sessions for both the painful and the asymptomatic hips. Considering the averaged values, the four tests that most commonly reproduced pain in the symptomatic hips were the quadrant, full flexion internal rotation (FFIR), flexion adduction internal rotation (FADDIR) and compression in adduction at 90° flexion (FADC). As expected, the prevalence of pain when testing asymptomatic hips was much lower than when testing symptomatic hips. However, the quadrant, FFIR and FADDIR were still amongst the top four most provocative tests for the asymptomatic hips. Tests that included internal rotation were much more likely to cause pain in symptomatic hips than those that included external rotation. Whilst most resisted and ‘functional’ tests reproduced pain in this group, the prevalence of a positive test was generally less than 10%. The majority of tests (18/30) performed on asymptomatic hips were pain free. Of those that were painful, the quadrant, FFIR and FADDIR tests were the most provocative.

Table 3.12 Prevalence of positive test responses

Test	Prevalence of +ve Test in Symptomatic Hip			Average over three sessions	
	Trial 1 (%)	Trial 2 (%)	Trial 3 (%)	Symptomatic Hip (%)	Asymptomatic Hip (%)
<b>Impingement Tests</b>					
Quadrant	83	83	72	80	23
FADDIR	61	72	67	67	13
FADC	78	72	50	67	6
FIRC	39	50	39	43	6
<b>Miscellaneous Tests</b>					
FF	59	65	59	61	4
FABER	56	61	50	56	2
BKFO	22	28	28	26	4
EPr	6	6	0	4	0
Log Roll	6	6	6	6	0
<b>Internal Rotation Tests</b>					
FFIR	89	78	68	78	10
FIR	56	50	50	52	10
IRSit	39	33	28	33	0
IRPr	61	56	50	56	4
IRSt	50	39	28	39	0
<b>External Rotation Tests</b>					
FFER	33	17	22	24	2
FER	26	17	28	24	6
ERSit	6	11	6	7	0
ERPr	6	6	6	6	0
ERSt	11	6	11	9	0
<b>'Functional' tests</b>					
ADDSt	33	44	11	30	0
ADDKn	50	50	44	48	0
FKn	33	22	17	24	0
Squat	6	0	6	4	0
Rise	0	0	0	0	0
<b>Resisted Tests</b>					
RAD	11	6	11	9	0
RAB	17	0	6	7	0
RF	17	6	11	11	0
RE	0	0	0	0	0
RIR	22	11	6	13	0
RER	17	6	6	9	0

FADDIR, Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FIRC, Flexion Internal Rotation Compression; FF, Full Flexion; FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; EPr, Extension in Prone; FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRPr, Internal Rotation in Prone; IRSt, Internal Rotation in Standing; FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERPr, External Rotation in Prone; ERSt, External Rotation in Standing; ADDSt, Adduction in Standing; ADDKn, Adduction in 4-point kneel; FKn, Hip flexion in 4-point kneel; Squat, squat to chair; Rise, rise from chair; RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°

### 3.5 Discussion

This study investigated volunteers with unilateral pain in the groin or buttock, recruited either by referral from a sports physician or through advertisements placed in a university setting. The participants were relatively young, engaged in society and able to participate in normal activities of daily living. However, most were unable to play their normal sports because of their hip pain. None had severe pain. The following discussion needs to be considered bearing in mind these characteristics of this cohort.

#### 3.5.1 Reliability of reports of pain reproduction

Whilst there is no level of reliability that has universal acceptance as being the minimum required before a physical test should be included in a clinical examination, there appears to be some consensus that a kappa, PABAK or ICC value of  $>0.6$  is required (Cadogan, Laslett, Hing, McNair, & Williams, 2011; Chinn, 1991; Cibere et al., 2008; Laupacis, Sekar, & Stiell, 1997; Martin & Sekiya, 2008). Cadogan et al. argue that this level of concordance should be accompanied by percent agreement in excess of 80% before it is appropriate to use such tests clinically. In respect to the within-session reliability of the reproduction of familiar pain, given the above parameters, 15 of the tests included in the current study are appropriate for clinical use (quadrant, log roll, squat, rise, BKFO, EPr, IRPr, ERSit, ERSt, ERPr, ADDKn, RAB, RAD, RF, RE). Ten tests demonstrated both reliability scores below 0.6 and percent agreement below 80%. For between-session reliability, nineteen tests demonstrated kappa above 0.6 and percent agreement above 80% (quadrant, log roll, squat, FADDIR, FABER, EPr, IRPr, IRSt, FFER, FER, ERSit, ERSt, ERPr, FKKn, RAB, RAD, RF, RIR, RER). Six tests demonstrated both reliability and percent agreement below these figures. Eleven tests met the criteria for inclusion in the clinical examination for both within and between-session reliability (quadrant, log roll, squat, EPr, IRPr, ERSit, ERSt, ERPr, RAB, RAD, RF).

We cannot compare these findings to other studies, as it appears that no previous investigation of the *intra*-examiner reliability of categorical decision-making regarding the presence or absence of pain provoked by physical tests of symptomatic hips has been conducted. Some studies (Cibere et al., 2008; Martin & Sekiya, 2008; Ratzlaff et al., 2013) have investigated the *inter*-examiner reliability of such decision-making. However, a comparison of *intra*-examiner reliability to *inter*-examiner reliability,

especially in light of differences in the definition of a positive test result, the experience and background of the examiners as well as the characteristics of the included participants across these studies, is perhaps not particularly meaningful or appropriate.

The results of the current study demonstrate that there is some inconsistency in the participants' responses when they are asked if a test reproduces their usual (familiar) pain or not. Bearing in mind that many of these tests provoke pain in asymptomatic hips, it may be that the requirement to commit to a decision that the pain provoked by a test was actually their 'familiar' pain was difficult. This would not be the case when it was obvious that their pain had been reproduced. However, if the pain was less severe or slightly different to their 'familiar' pain, this may have created some indecision. Whilst the method of application was standardised, it is also possible that there was some variation in the positioning of the limb during testing or the amount of load applied by the examiner. Either of these possibilities could have affected the reliability estimates for these tests. The influence of these factors could have been reduced by the use of positioning devices and a force transducer (to measure load at the end of the range) for each test. However, these options were discarded a priori as they were considered not to be pragmatic for either the clinical application of these tests or for the diagnostic accuracy study that follows the current study (Kottner et al., 2011; Stockendahl et al., 2006).

### **3.5.2 Reports of pain intensity**

Fourteen tests included in the current study demonstrated substantial (0.61 to 0.80) to almost perfect (0.81 to 1) within-session reliability for ratings of pain intensity (Landis & Koch, 1977). Seventeen tests showed this level of reliability for between-session testing. Twelve tests (quadrant, FADDIR, FADC, FF, BKFO, FFIR, FIR, IRSit, IRPr, IRSt, FK<sub>n</sub>, RF) fell within this range for both within and between-session reliability. Overall, 19 of the 30 tests examined demonstrated reliability scores higher than the 0.6 level argued to be appropriate for inclusion in a clinical examination (Cibere et al., 2008; Laupacis et al., 1997; Martin & Sekiya, 2008).

The only other study (Cliborne et al., 2004) that has examined the reliability of patient reports of pain intensity during the application of physical tests of the hip, reported excellent reliability for FABER, hip flexion, scour and a 'functional squat' tests. However, these researchers asked patients with symptomatic *knee* arthritis to rate pain intensity during the application of *hip* joint tests. In the absence of hip pain or

pathology, it is difficult to interpret these findings. Also, the test-retest period in this study was only 2-minutes compared to 60-minutes in the current study. Short intervals between test sessions increase the likelihood of the participant remembering their previous scores and will therefore enhance reliability estimates (DeVon et al., 2007). These factors make it difficult to make any comparison between the study by Cliborne and colleagues and our own.

It is interesting to note that most resisted tests and most tests that included external rotation demonstrated poor reliability. These tests were not particularly provocative in the group of patients included in this study. The mean pain intensity for most of these tests was below 1 on the NPRS for all trials. Conversely, the tests with the highest reliability typically demonstrated the highest prevalence of a positive test and highest mean pain intensity scores. When pain intensity is low, small changes in intensity will result in large differences in terms of the percentage of the original rating. This will have an adverse effect on the estimates of reliability for tests that cause little pain. This finding has implications for both the clinical and research settings. Where important diagnostic or therapeutic decisions are made on the basis of patient ratings of pain intensity, it would be prudent to consider findings of the tests that are more provocative than those that cause pain of very low intensity

This point is reinforced when the error associated with measurements of pain intensity and the minimal clinically important difference (MCID) for the NPRS are considered. SEM values across the tests included in our study ranged from 0.2 to 1.8 points on the NPRS. The average SEM was 0.8 points for both within and between-session situations. A SEM of this size means that a patient may rate pain as 0.8 points above or below the 'true' value at a given point in time. In the 'worst case' scenario, a patient might report pain intensity, for example, of 7 one occasion, 7.8 the next and 6.2 on a third occasion, even though the actual pain intensity has remained at 7 over this period of time. The difference between these two scores is 1.6 points. This value is very similar to the 2-point shift reported as the MCID for changes in pain intensity when using the NPRS for patients with musculoskeletal pain (Childs et al., 2005; Farrar et al., 2001).

Based on our results, patients who report low baseline pain intensity (say <2 points) for a given test are less likely to provide a reliable measure of pain intensity during the reapplication of that test at a subsequent time. A clinically important change in intensity of this pain would require a total abolition of the pain. A 50%, or even 80% reduction,

would not be convincing given the possibility that the original rating could not be relied upon to be 'accurate' and that the shift is less than 2 points on the NPRS.

### **3.5.3 Prevalence of painful responses**

The current study provides information regarding the prevalence of reports of pain during the application of physical tests for the hip. Whilst the study was not powered to determine population prevalence, the prevalence within the study cohort was of interest. Many of the tests included in this study caused pain in both the symptomatic and asymptomatic hips. The tests that most frequently reproduced pain in symptomatic hips were very similar to those that were provocative in asymptomatic hips. There were only subtle changes in the ranking (in terms of highest to lowest prevalence of pain production) for these tests between the symptomatic and asymptomatic sides. Although the ranking was similar, the actual prevalence of a positive test result was significantly higher in symptomatic hips. This is intuitive in that one would expect a painful hip to be more likely to hurt with mechanical loading than a 'normal' hip. Also, tests that demonstrated a relatively low prevalence of pain reproduction in symptomatic hips (e.g. resisted tests with an average prevalence of 14%) were rarely provocative in participants with asymptomatic hips.

The prevalence of painful responses in the asymptomatic hips was surprisingly high for a number of tests. The quadrant test provoked pain in 23% of asymptomatic hips. The FADDIR, FIR and FFIR tests were painful in around 10% of all asymptomatic hips. The only other study to investigate prevalence of positive tests in asymptomatic hips is Prather et al. (2010) who reported a prevalence of 7.1% for FADDIR, 3.6% for FABER and 0% for log roll. These figures are congruous with those of the current study. The mean age, age range and BMI of the participants in this study were remarkably similar to our own.

It is likely that the pain associated with testing asymptomatic hips is due to stress on normal tissues that are placed under end range stretch or compressive load during the test. However, the findings of the current study should be considered alongside research that has demonstrated pathological/structural changes, identified by medical imaging techniques of the hip (including, x-ray, ultrasound and MRI) in people who do not have any history of pain or functional disability (Frank et al., 2015; Hack, Di Primio, Rakhra, & Beaulé, 2010; Jung et al., 2011; Kumar et al., 2013; Lanyon, Muir, Doherty, & Doherty, 2003; Lee, Armour, Thind, Coates, & Kang, 2015; Register et al., 2012). It is

clear that people may have significant degenerative and structural changes that do not cause symptoms on a day-to-day basis. Our study also demonstrates that the prevalence of positive tests for the hip was much higher for tests that incorporate internal rotation and/or flexion of the hip compared to those that include external rotation and/or less than 90 degrees of flexion, regardless of whether the hip was symptomatic or not. Flexion and internal rotation are the key components of the various ‘impingement’ tests used on the hip. Such tests are proposed to engage the femoral head-neck junction against the labrum and acetabular rim. Consequently, these tests are likely to load the structures commonly identified as being abnormal in asymptomatic populations (as described in the abovementioned studies). It may be that the tests used in the current study were provocative in asymptomatic hips because of the presence of underlying structural abnormalities that were not severe enough to cause symptoms with normal activities of daily living.

A key clinical implication of these findings is that clinicians should be cognisant of the fact that if physical tests are provocative in people without hip pain, then a positive test in a patient with hip pain does not necessarily implicate the hip joint. Many ‘hip’ tests have the potential to load other structures (e.g. the sacroiliac joint, the lumbar spine, the pubic symphysis) and false positive tests can arise when a test is not specific to the structure targeted with the test.

### **3.6 Limitations**

This study explored the reliability of reports of pain provocation and of pain intensity obtained by an individual examiner. The investigation of intra-examiner reliability was a purposeful decision designed to replicate the common clinical scenario where assessment is undertaken by one clinician before and after an intervention and at subsequent treatment sessions. Inter-examiner reliability was not examined and these results should not be extrapolated to situations where different clinicians are assessing these factors.

The amount of force applied at the end of range for each of the tests was not standardised in an objective manner. Pain intensity typically increases as increasing force is applied during the application of physical tests and the use of a force transducer to ensure that the same degree of load was applied for each test and each session was considered. However, standardising the level of force would increase the likelihood of false negative results in patients that required more force than the ‘chosen standard’ to

provoke their pain. Similarly, patients with pain that was easily provoked may have been harmed if the standardised force was more than what was appropriate to apply to their injury. Furthermore, the intent of this study was to establish the reliability of patient reports of pain intensity during the application of tests performed in a manner that reflects clinical practice. Consequently, the examiner applied sufficient force to determine end range of movement (as determined by significant tissue resistance) or until the participant's pain response indicated that the application of further load would be inappropriate. It may be argued that the examiner in the current study modified the force applied in a manner to try to achieve the same level of pain intensity as that produced in earlier sessions. However, this is unlikely given the number tests performed (n=30), the fact that they were performed in random order, that they were performed one hour and then days apart and that scores from each session were secured from the examiners view immediately after each session.

### **3.7 Conclusion and implications**

This study has demonstrated that patient reports of pain reproduction and ratings of pain intensity during the application of PPT's to the hip, for the most part, have sufficient reliability to justify their use in both the clinical and research environments. Whilst some tests demonstrated better reliability than others, this appears to relate to how provocative that test was i.e. tests that reproduced pain of high intensity were more reliable than those that caused minimal or low intensity pain. Hence, we suggest that when the assessment of the effect of an intervention *on pain* is required, that tests that create pain of a 'higher' intensity should be used as baseline measures. Assessment of pre to post treatment changes in pain intensity are likely to be more accurate when such tests are used than when tests that create little pain are employed.

These findings also have implications for this thesis. Changes in pain intensity following the intra-articular injection of anaesthetic will be used as the reference standard in the diagnostic accuracy study that is the key focus of the thesis. The current study has provided evidence that ratings of pain intensity can be relied upon to give a true reflection of the effect of the anaesthetic provided the following criteria are met:

- that the pain produced by a test must be the same pain for which they have sought treatment (a familiar pain)
- only tests that reproduce pain of an intensity of 2 or more points on the NPRS should be considered as a positive test

- a reduction of pain intensity of less than 2 points on the NPRS should not be considered significant or taken into account when calculating response to the anaesthetic.

## **Chapter 4 Measurements of Strength and Range of Movement in Painful and Non-Painful Hips**

This Chapter relates specifically to Questions 2 and 3 of this thesis:

*What is the between-session intra-examiner reliability of measures of strength for both the symptomatic and asymptomatic hip in people with unilateral hip pain?*

*What is the within and between-session intra-examiner reliability of measures of range of movement for both the symptomatic and asymptomatic hip in people with unilateral hip pain?*

### **4.1 Introduction and background**

Hip muscles have an important role in stability and function of the hip joint (Neumann, 2010; Retchford, Crossley, Grimaldi, Kemp, & Cowan, 2013; Zifchock, Davis, Higginson, McCaw, & Royer, 2008). Decreased lower-extremity muscle strength has been shown to be one of the strongest predictors of falls in older adults (Moreland, Richardson, Goldsmith, & Clase, 2004). There is evidence that decreased strength and decreased range of movement are associated with painful hip pathology (Agricola et al., 2013; Arokoski et al., 2002; Harris-Hayes et al., 2014; Judd, Thomas, Dayton, & Stevens-Lapsley, 2014; Kemp et al., 2014b). Impairment in muscle function can affect optimization of the load on the hip joint and its surrounding structures during movement, resulting in increased contact forces that can contribute to the development and progression of pathology (Bergmann, Graichen, & Rohlmann, 2004). Some authors have proposed that strength differences between sides and/or between agonist-antagonist muscle groups may contribute to injury or re-injury (Knapik, Bauman, Jones, Harris, & Vaughan, 1991; Tyler, Nicholas, Campbell, & McHugh, 2001; Yeung, Suen, & Yeung, 2009).

Strength testing appears to be useful from a diagnostic perspective, with studies demonstrating that differences in strength can discriminate patients from controls and various pathologies from one another. For example, Gruther et al. (2009) demonstrated that patients with non-specific chronic low back pain have significantly weaker back extensors and flexors than normal controls. Jeon, Chung, Lee, Son, and Kim (2013) reported that testing for concomitant weakness of the hip abductors differentiates peroneal neuropathy from lumbar radiculopathy, as a cause of a foot drop. These

authors reported that hip abductor weakness was 86% sensitive and 96% specific for this purpose. With respect to the hip, Arokoski et al. 2002 reported that men with hip OA have significantly lower strength of the abductors, adductors and flexors than normal controls. Similarly, Harris-Hayes et al. (2014) demonstrated that people with chronic hip joint pain have significant hip muscle weakness compared to controls.

The identification of changes in range of movement is also of diagnostic value, with a loss of internal rotation and/or hip flexion being two key criteria for a clinical diagnosis of osteoarthritis of the hip (Altman et al., 1991; Birrell et al., 2001). A loss of hip extension was shown by Joe et al. (2002) to have a high specificity (92%), but a low sensitivity (19%), with respect to the identification of avascular necrosis of the femoral head in asymptomatic HIV-infected patients. In contrast, increased internal rotation has been associated with the development of lower limb injury (Wyss, Clark, Weishaupt, & Nötzli, 2007; Zifchock et al., 2008).

This thesis focuses on determining the diagnostic accuracy of the clinical examination of the hip. There appears to be sufficient evidence to suggest that the identification of changes in strength and ROM may be an important component of the hip examination and that further exploration of the diagnostic accuracy of such changes is warranted. However, prior to their inclusion in the primary study of this thesis, it was considered important that measurements of strength and ROM made by this researcher (the PhD candidate) were reliable. Consequently, the current study was designed to address Questions 2 and 3 of this thesis.

A decision was made to use a previously validated hand held dynamometer (HHD) that incorporates both a force transducer (to measure strength) and gravity dependent inclinometer (to measure range of movement) (Industrial Research Ltd; Christchurch; NZ). The decision to use a hand held device rather than more sophisticated isokinetic testing was essentially for pragmatic reasons. The main study in this thesis needed to be conducted onsite at a private radiological clinic so that participants could be assessed immediately before and after the fluoroscopy-guided anaesthetic injection used as the reference standard. It would be impractical to use an isokinetic device in this setting.

This chapter begins with a literature review that first considers previous research that has examined the *reliability* of measures of hip ROM and strength. Next, literature that has investigated hip strength and/or ROM *differences* is presented. A narrative review

of research that considers the key methodological factors important to the design and conduct of reliability studies is also provided. Finally, the chapter presents the methods, results and discussion for the current study.

## 4.2 Literature review

An initial literature search indicated that there were a limited number of studies that had investigated differences in hip strength and ROM between sides in people with unilateral hip pain. Therefore, studies that compared such measures in people with symptomatic hips to people with ‘normal’ hips, and studies that considered side-to-side differences in people with ‘normal’ hips were included. With respect to *reliability* of measurements of strength, given our intent to use a HHD for such measures, only research that used such a device to measure hip strength were included in this section of review. However, studies that employed other measurement tools (e.g. strain gauge dynamometers and isokinetic devices) were included in the review of evidence that has investigated side-to-side differences in strength. For ROM, due to a relatively small number of such studies that have investigated the intra-examiner reliability of an inclinometer, we also included studies that employed a goniometer to measure hip ROM.

This review was also performed to identify relevant research that informed the design and methods of the study reported in this chapter. The specific aims of this review were to answer the following questions:

- What is the current evidence regarding the intra-examiner reliability of measures of strength and range of movement of the hip joint in people with hip pain?
- What evidence exists regarding the diagnostic utility of side-to-side differences in strength of hip muscles in people with unilateral hip joint pain?
- What evidence exists regarding differences in hip strength in people with hip pathology versus those with normal hips?
- What evidence exists regarding the diagnostic utility of side-to-side differences in range of movement in people with unilateral hip pain?

The initial search was performed (using the search strategy detailed in Chapter 2) prior to the commencement of data collection for the current study (in May 2010). It was

updated with a follow-up search performed in March 2015. Key concepts were identified and searched separately in 5 main categories summarised as: 1) "hip joint" OR "hip pain" OR "groin pain" OR groin OR hip 2) reliability OR accuracy OR consistency OR validity 3) strength OR "muscle strength" OR "peak force" OR force 4) ROM OR "range of movement" OR "range of motion" range 5) "Physical examination" OR "tests" OR "clinical examination" OR "objective examination" OR "impairment".

Where relevant high quality systematic reviews were identified, their findings are first presented. This is followed by any relevant experimental studies published since these reviews. Results are detailed below under sub-headings that reflect the review aims. In total, 39 relevant studies were identified. Twenty-three of these investigated reliability of strength measures in the hip using a hand-held dynamometer, 14 studies investigated reliability of range of movement measures in the hip and 3 studies investigated both strength and ROM. Three relevant systematic reviews (Diamond et al., 2015; Dobson, Choi, Hall, & Hinman, 2012; Loureiro, Mills, & Barrett, 2013) and one narrative review were identified (Bohannon, 2012).

#### **4.2.1 Reliability of measures of range of movement and strength**

The systematic review by Dobson et al. (2012) included 15 studies, 6 of which measured strength and 9 measured ROM. All studies included participants with hip pathology (including OA, fracture and 'groin pain'). Eleven studies investigated reliability of these measures, three investigated validity and one considered both reliability and validity. The remaining study investigated internal consistency of strength measures. The quality of the Dobson et al. systematic review was evaluated using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Liberati et al., 2009; Moher, Liberati, Tetzlaff, & Altman, 2009). Whilst these guidelines were developed primarily to help ensure clarity and transparency in the reporting of systematic reviews, an appreciation of the quality of a review can be ascertained by considering each item in the guidelines against the content provided by the authors of a systematic review (Liberati et al., 2009; Moher et al., 2009). Dobson et al. provided an appropriate rationale and clear objectives for their review along with detail of their eligibility criteria and sources of information. Comprehensive details of the search strategy, study selection and data extraction processes were included. These authors used the previously validated 'Consensus-Based Standards for the Selection of Health Measurement Instruments' (COSMIN) critical appraisal tool developed for evaluating the methodological quality of studies included in

their review (Mokkink et al., 2010; Terwee et al., 2012). Findings of the individual studies were clearly presented, as were the details of the assessment of risk of bias. Hence, this review provided sufficient detail to allow the quality of the review to be evaluated. On the basis of this information, the systematic review by Dobson et al. (2012) appears to be of a high standard and consequently their conclusions are likely to represent 'best-evidence' at this point in time.

This review included four studies (Holmich, Holmich, & Bjerg, 2004; Malliaras, Hogan, Nawrocki, Crossley, & Schache, 2009; Pua, Wrigley, Cowan, & Bennell, 2008; Sherrington & Lord, 2005) that have investigated the intra-examiner reliability of strength measures in people with hip pain. Dobson et al. reported that only two of these studies (Pua et al., 2008; Sherrington & Lord, 2005) were good quality. Both studies measured isometric strength with a HHD. Pua et al. measured strength of internal and external rotators, flexors, extensors and abductors in 22 patients (mean age  $62 \pm 8.9$  years) with symptomatic osteoarthritis of the hip. They reported intraclass correlation coefficients (ICC's) between 0.84 (95% CIs 0.55, 0.94) and 0.98 (95% CIs 0.94, 0.99). Sherrington and Lord measured strength of flexors and abductors in 30 patients (mean age  $79 \pm 10$  years) recovering from a hip fracture (1 week to 57 weeks post fracture). These authors reported ICC's of 0.80 (95% CIs 0.61, 0.91) for flexion and 0.86 (95% CIs 0.71, 0.94) for abduction. On the basis of these studies, Dobson et al. concluded that a HHD could be recommended for measuring the strength of hip muscles.

Dobson et al. also included 3 studies that had examined between-session reliability measures of ROM. Dobson et al. considered that only two were good quality studies (Holm et al., 2000; Pua et al., 2008). Pua and colleagues investigated the *intra*-examiner reliability of measures of ROM for internal and external rotation, flexion and extension using an inclinometer. They reported ICC's between 0.86 (95% CIs 0.67, 0.94) and 0.97 (95% CIs 0.93, 0.99). Holm et al. investigated the *inter*-examiner, between-session reliability of ROM measures using a goniometer in 25 patients with symptomatic OA. These authors reported ICC's ranging from 0.50 (no CIs reported) for adduction to 0.94 for extension. In this study, pooled results from separate 'teams' of examiners were presented. These teams included one 'team' that was an orthopaedic surgeon estimating ROM visually, one team that had 2 novice physiotherapists, and another that had two expert physiotherapists. These pooled results make it difficult to interpret the findings from this study. However, on the basis of these two studies, Dobson and colleagues concluded that measurements of hip ROM (using either an inclinometer or goniometer)

in people with hip pain could be recommended for flexion, extension, internal/external rotation and abduction.

A number of studies identified by the current search were not included in the Dobson et al. review. However, the majority of these investigated reliability in people *without* hip pain. Two additional studies (Klassbo, Harms-Ringdahl, & Larsson, 2003; Nussbaumer et al., 2010) that investigated ROM in patients with hip pain and one study (Arnold, Warkentin, Chilibeck, & Magnus, 2010) that investigated strength were identified. The earliest of these studies (Klassbo et al.) was primarily a study designed to explore capsular patterns of the hip joint. Within this publication, Klassbo et al. reported that they had performed a separate reliability study of ROM measures obtained with a goniometer. The study included 14 people with hip pain and 6 'normals'. Unfortunately, only brief details regarding the conduct of this study were provided making it difficult to determine its overall quality. They reported point estimates of ICC's between 0.56 (for extension) to 0.92 (for flexion) without any confidence intervals. Given the lack of detail, these results should be considered with a degree of caution.

Nussbaumer et al. investigated ROM using a goniometer and an electromagnetic tracking system (ETS) in 15 patients with femoroacetabular impingement (FAI) and 15 matched, healthy controls. Whilst these authors reported excellent test-retest reliability for both devices (ICC's above 0.84 for all ROM assessments), they also reported that the two devices provided different measurements. Goniometric measurements were generally greater than those obtained by the ETS. Agreement between these instruments was poor for flexion, adduction and external rotation (ICC of 0.44, 0.53 and 0.54 respectively). In contrast, agreement was excellent for both abduction (ICC =0.93) and internal rotation (ICC=0.87).

Arnold et al. investigated the within-session (same day) and between-session (a day apart), *intra-examiner* reliability of strength measures using a HHD in 18 patients (11 with symptomatic hip or knee OA). They measured hip flexion, extension and abduction bilaterally. Hip flexion was measured in both sitting and standing, and abduction in both supine and standing. The authors did not report values for individual tests for within-session measurements. Instead, they stated that they were 'high' and that values ranged from 0.90 to 0.98 (no confidence intervals reported). However, they did report values for the between-session reliability. These results demonstrated that test position did not significantly alter the reliability of these measures with ICC's between

0.84 and 0.94. Perhaps the most interesting finding in this study was that reliability for the right hip was consistently lower than for the left. Arnold et al. did not comment on this finding.

### **Summary**

In summary, two studies (Pua et al., 2008; Sherrington & Lord, 2005) have demonstrated excellent between-session, intra-examiner reliability of strength measurements using a HHD in people with hip pathology. One study (Arnold et al., 2010) has investigated the within-session reliability of such measures, although not all participants in this study had painful hips. Only Pua et al. has investigated the reliability of measures of ROM obtained with inclinometer in people with hip pain, reporting excellent reliability. Although this evidence suggests that measurements of strength with a HHD and ROM with an inclinometer are reliable, there is insufficient evidence to be confident that this will be the case for this examiner, in patients with hip pain, in the circumstances of the proposed diagnostic accuracy study. Hence, it would be prudent to conduct a study that demonstrates that this is the case.

#### **4.2.2 Side-to-side differences in strength in people with unilateral hip pathology**

Loureiro et al. (2013) performed a systematic review of evidence investigating muscle weakness in people with hip osteoarthritis. The authors provided an appropriate rationale and clear objectives along with detail of eligibility criteria and sources of information. Comprehensive details of the search strategy, study selection and data extraction processes were included along with an explanatory flow diagram. The methodological quality of included studies was assessed using the Cochrane Collaboration's bias assessment tool (Higgins et al., 2011) and risk assessment details were provided for each study. Results of the individual studies were clearly presented in various tables and associated text. These authors reported that all of the studies included in their review had either a low or moderate risk of bias providing some confidence in the integrity of their findings. Hence, this review provided sufficient detail to allow the quality of the review to be evaluated. On the basis of this information, this systematic review appears to be of a high standard and consequently their conclusions are likely to represent 'best-evidence' at this point in time.

This review included 3 relevant studies (Arokoski et al., 2002; Rasch, Byström, Dalen, & Berg, 2007; Rasch, Dalén, & Berg, 2010) that compared side-to-side strength in

people with unilateral hip OA. Rasch et al. (2007) measured side-to-side peak isometric force (with a strain gauge dynamometer) in 22 elderly (mean age  $67 \pm 7$  years) patients with unilateral OA. All participants were on a waiting list for a hip joint replacement. Hip abduction, adduction, flexion and extension were measured with the hip positioned in  $45^\circ$  of hip flexion and the patient standing. These authors reported statistically significant reductions in strength on the painful side, ranging from 11% (for adduction) to 27% (for flexion). This research group reassessed strength at 6 and 24 months post hip joint replacement. In a follow-up study (Rasch et al., 2010) they reported that at 24 months, the hip abductors were still 15% weaker on the painful side whereas the extensors, flexors and adductors were not statistically different between sides.

In contrast to the studies by Rasch and colleagues, the study by Arokoski et al. included patients with moderate OA (not end-stage). In this study, 15 patients with unilateral hip OA and 12 patients with bilateral hip OA were compared to 30 aged-matched, normal controls (mean age  $56.3 \pm 4.5$  years). The authors compared side-to-side peak isometric torque of hip abduction and adduction (measured with a strain gauge dynamometer) and hip flexion and extension (measured with an isokinetic dynamometer) with the hip positioned in zero degrees of flexion. For side-to-side comparison in patients with unilateral OA, strength values of the symptomatic hip were compared to the asymptomatic hip. For patients with bilateral OA, the 'better and worse' hips were identified by the severity of OA present on radiographs using the Kellgren-Lawrence (Kellgren & Lawrence, 1957) grading criteria. The hip that had 'more severe radiographic changes' was considered the 'worse' hip and was compared to the 'better' hip. Data from the symptomatic hip and the 'worse' hip were combined and compared as a group to the combined data from the asymptomatic or 'better' hip. Arokoski et al. reported that the 'worse' hip group was significantly weaker in both flexion (21%) and extension (22%) than the 'better' hip group. Arokoski et al. reported that there was not any correlation with pain intensity and strength values, suggesting that the weakness was not due to pain inhibition. In contrast to this, no significant side-to-side differences in adduction or abduction strength were seen, despite the fact that these muscles demonstrated significantly reduced cross-sectional area (CSA) on MRI. This led these researchers to conclude that a reduction in CSA in patients with hip OA is not a direct indicator of decreased hip strength. Arokoski et al. also investigated side-to-side strength in the controls, reporting that the only significant difference they observed was that isometric extension strength was 15.8% stronger on the right side. On the basis of

the evidence presented in these three studies, Loureiro et al. (2013) concluded that there was 'strong evidence' of weakness of the muscles of the affected hip in people with OA.

The current search identified a number of studies that have compared side-to-side strength in people with *normal* hips (Bandinelli et al., 1999; Bohannon, Vigneault, & Rizzo, 2008; Cichanowski, Schmitt, Johnson, & Niemuth, 2007; Jacobs, Uhl, Seeley, Sterling, & Goodrich, 2005; Kemp, Schache, Makdissi, Sims, & Crossley, 2013; Niemuth, Johnson, Myers, & Thieman, 2005; Phillips, Lo, & Mastaglia, 2000; Rasch, Dalen, & Berg, 2005; Thorborg, Coupe, Petersen, Magnusson, & Holmich, 2011a; Thorborg et al., 2011b). Findings from these studies are conflicting, with some authors reporting no significant differences between sides (Bandinelli et al., 1999; Bohannon et al., 2008; Cichanowski et al., 2007; Niemuth et al., 2005; Rasch et al., 2005; Thorborg et al., 2011a), whilst others report that such differences do exist for some muscle groups (Bohannon et al., 2008; Jacobs et al., 2005; Phillips et al., 2000; Thorborg et al., 2011a; Thorborg et al., 2011b). A detailed examination of this literature is beyond the scope of this thesis.

### **Summary**

In summary, there is some evidence that the painful hip in people with severe, unilateral OA is weaker than their contralateral side. Also, one well conducted study (Arokoski et al., 2002) suggests that this is the case for people with less severe OA. However, this evidence is not compelling and it seems that it would be appropriate to further investigate side-to-side strength differences in patients with hip pain. The inclusion of patients with hip pain associated with a wider range of pathologies (e.g. labral tears, FAI) might be informative. OA is typically a chronic, progressive condition where the likelihood of associated disuse atrophy of the muscles of the symptomatic hip is high. Similarly, the intra-articular swelling and inflammation associated with OA may lead to neural inhibition and consequential muscle weakness (Rice & McNair, 2010). Hence, differences in side-to-side strength may be more likely in such patients than those with less chronic conditions. If significant differences in side-to-side strength exist in a broader range of hip pain patients, exploration of the diagnostic value of such findings would be warranted.

### 4.2.3 Comparisons of strength between symptomatic hips and normal controls

Comparisons of the symptomatic hip in patients with hip joint pain to asymptomatic matched controls have also been made (Arokoski et al., 2002; Casartelli et al., 2011; Harris-Hayes et al., 2014; Kemp et al., 2014b; Klausmeier, Lugade, Jewett, Collis, & Chou, 2010; Rasch et al., 2005). As previously described (see page 89), Arokoski et al. compared patients with hip OA to aged-matched, normal controls. They reported that the adductors, abductors and flexors were weaker (25%, 31% and 18% respectively) in the OA group when compared to the controls. No significant difference in the strength of the extensors was observed between these groups.

Rasch et al. (2005) compared peak isometric force (measured in Newtons) of 10 'young' (age  $36 \pm 6$  years) and 13 'elderly' (age  $69 \pm 8$  years), healthy volunteers to 11 patients (age  $69 \pm 8$ ) with unilateral OA. These authors reported statistically significant weakness in the patients compared to the controls for hip extension, flexion and abduction (24%, 27% and 32% weaker respectively). Rasch and colleagues also reported that there was not any difference between left and right sides in the healthy subjects included in this study. Whilst this comparison of force between sides for an individual is appropriate, it is not so for comparisons between subjects. The length of the lever arm will influence measured force values. With different subjects, the lever arm is likely to vary. Hence, torque is a more appropriate unit of measurement. Consequently, the findings of Rasch et al. in respect to decreased force in people with OA versus controls need to be considered with caution. However, similar strength differences between controls and people with hip pain were reported by Klausmeier et al. (2010). In this study, isometric hip abductor strength was measured with a KIN-COM dynamometer in 23 patients (mean age 57) who were scheduled for a total hip arthroplasty (THA) and compared to 10 healthy, age-matched controls. A statistically significant difference in torque (calculated by multiplying force values by the length of the moment arm) was observed between groups, with the group mean for the surgical candidates being 30% lower than that of the control group (0.47 Nm/kg and 0.67 Nm/kg respectively).

This trend indicating weakness of hip muscle in people with hip pathology compared to controls is supported by the findings of Casartelli et al. (2011) who compared hip strength in 22 patients (mean age  $32 \pm 9$ ) with FAI to 22 age-matched controls. In this study, peak isometric force of hip abduction and adduction (with the participants in side

lying) and internal and external rotation (with the patients sitting and hips at  $90^{\circ}$  flexion) was measured using a HHD. Hip flexion and extension strength was measured using an isokinetic dynamometer (Biodex) and with the hip at  $45^{\circ}$  of flexion. These authors reported statistically significant reductions in isometric torque of the symptomatic hip in FAI patients compared to controls for adduction, abduction, flexion and external rotation (28%, 11%, 26% and 18% weaker respectively). However, no between group differences were seen for extension or internal rotation torque. Casartelli and colleagues also reported the mean pain intensity (measured with a 100mm visual analogue scale) experienced during the strength tests. Whilst no controls experienced pain during strength testing, pain intensity reported by the patients ranged from  $18 \pm 20$  mm (for resisted extension) to  $25 \pm 22$  mm (for resisted internal rotation). These authors suggested that this pain (and/or fear of pain) might have contributed to the weakness observed in the patient group. However, they also suggested that muscle atrophy might have been a factor.

The strength of the symptomatic hip in 84 patients (mean age 36 years) with chondrolabral pathology (identified at arthroscopy) was compared to that of 60 healthy age-matched controls by Kemp et al. (2014b). Isometric peak force of hip abduction, adduction, extension, flexion and both rotations were measured using a HHD. Hip flexion strength was measured with the participants in sitting whilst all other measures were performed with the hip in zero degrees of flexion/extension and zero degrees of abduction/adduction. Torque was calculated by multiplying force values by the length of the moment arm. These authors observed significantly reduced torque in the patient group for all muscle groups except internal rotation. Unfortunately, only the mean difference in normalized scores (Nm/kg) between groups was reported, making it difficult to compare percent differences with other studies.

The most recent relevant study (Harris-Hayes et al., 2014) compared the strength of hip abductors and internal and external rotators in 35 patients (mean age  $28.2 \pm 5$  years) with chronic hip joint pain (CHJP) to 35 asymptomatic age-matched controls. Peak isometric force was measured using a HHD. Hip rotation was measured with the participants sitting with their hip flexed  $90^{\circ}$  and fully internally rotated for the internal rotation test and fully externally rotated for the external rotation test. Strength testing of the rotators was repeated with the participant supine (hip in zero degrees of flexion) and the relevant muscle group in a shortened position. Abduction strength was measured in side lying with the hip abducted  $15^{\circ}$  (and zero degrees of flexion and rotation). Torque

was calculated and then normalised to weight and height. The findings of this study are similar to the previous studies in that people with CHJP were significantly weaker (from 16% to 28%) than the controls. In contrast to the previous studies, these authors observed significant weakness (28%) of the internal rotators. This finding may be related to the position of testing adopted by these authors. This is the only study, identified in the current search, that has tested the rotator muscles in an inner range position, a position that could have compromised the muscles ability to generate a maximum voluntary contraction (Gordon, Huxley, & Julian, 1966; Ward, Winters, & Blemker, 2010). Harris-Hayes et al. also reported that the *asymptomatic* hip in the participants with CHJP was weaker than the controls (external rotators 18% and abductors 16% weaker). This may be of significance in future research and clinical practice when the asymptomatic hip is used as a ‘control’ for patients with unilateral hip pain.

### **Summary**

On the basis of the findings reported by these researchers, there appears to be a consensus that symptomatic hip pathology is associated with weakness of the hip muscles when compared to age-matched controls. Whilst there is some variance in the degree of weakness reported and of the muscles involved, the trend is clear. These variances most likely relate to characteristics of the participants included in the respective studies and to differences in methods employed across the studies. This evidence suggests that the identification of hip muscle weakness may help differentiate pathological hips from normal hips.

#### **4.2.4 Side-to-side differences in range of movement**

A recent systematic review (Diamond et al., 2015) of evidence investigating range of movement in people with femoroacetabular impingement compared to controls included 12 studies of relevance to the current study. This review was assessed following the PRISMA guidelines. The authors provided an appropriate rationale and clear objectives along with detail of their eligibility criteria and sources of information. Comprehensive details of the search strategy, study selection and data extraction processes were provided. Methodological quality of the included studies were scored as ‘moderate to high’ using the Newcastle-Ottawa Scale (a tool designed for cohort and case-control studies, which is reliable and valid for assessing the quality of non-randomized studies) (Wells et al., 2015). Results of the individual studies were clearly presented in various

tables and associated text. This review appears to be of a high standard and consequently their conclusions are likely to represent ‘best-evidence’ at this point in time. Most of the studies included in this review used 3-dimensional motion analysis or CT to evaluate ROM although one, Nussbaumer et al. (2010), used a goniometer. Based on the evidence presented in the studies included in their review, Diamond and colleagues concluded that individuals with symptomatic FAI have decreased flexion and internal rotation (at 90° flexion) on the symptomatic side and that this restriction is at least partially due to bony impingement.

Earlier studies (Boone & Azen, 1979; Roaas & Andersson, 1982) have reported no significant differences in range of movement between sides in people with *normal* hips. An interesting and more recent study (Larkin, van Holsbeeck, Koueiter, & Zaltz, 2015) used ultrasound to determine the ‘impingement-free’ range of movement, bilaterally, in 40 asymptomatic young males (mean age 28 years). These authors passively flexed the hip until a point where labral deflection was identified (i.e. the point that the labrum was being impinged) and then continued through the range until there was bony abutment of the femoral head-neck against the acetabular rim. ROM was measured with a goniometer at each of these two points. Consistent with previous research, Larkin and colleagues reported that there were no differences between left and right sides in terms of ROM. They reported impingement free mean range for the left hip of  $68^{\circ} \pm 17^{\circ}$  and  $68^{\circ} \pm 16^{\circ}$  for the right hip. ROM determined by bony abutment was  $97^{\circ} \pm 6^{\circ}$  for the left side and  $96^{\circ} \pm 6^{\circ}$  for the right. The difference between impingement free ROM and bony abutment that they have highlighted may be of importance in people with symptomatic labral pathology.

### **4.3 Methodological considerations**

The following section summarises key considerations crucial to the design, conduct and analysis of the study presented in this chapter, with reference to the evidence identified in the current literature search.

#### **4.3.1 Instrumentation**

Whilst HHD’s and inclinometers can provide objective measures of strength and ROM, reliable and accurate results are dependent on key factors. Most importantly the instrument must be properly calibrated and have a high enough ‘ceiling’ to allow for any forces that they are required to measure. Also, the HHD must have adequate padding so that patients don’t feel discomfort during testing that might limit their ability

to generate peak forces (Bohannon, 2012). The reliability and validity of the HHD used in the current study has been investigated by Janssen and Le-Ngoc (2009). These authors reported that accuracy of this device in a laboratory setting was  $\pm 1^\circ$  for angle and  $\pm 1\text{N}$  for force and that intra-examiner reliability for peak torque and start and end ROM were excellent (ICC's 0.99, 0.98, and 0.99 respectively).

#### **4.3.2 Position and stabilisation of patient**

The ability of a muscle to generate force is influenced by the length at which that muscle is positioned (Gordon et al., 1966). During measurement of maximal isometric strength, the initial position of the limb determines the length of the muscle being tested. Whilst the rationale for choosing one particular test position over another is a consideration for investigators (Ward et al., 2010), changes from this position will alter the force generating capacity of the muscle (Bohannon, 2012; Ward et al., 2010). The key implication in respect to the reliability of strength testing is that there must be consistency in the initial positioning of the limb and careful attention paid to the maintenance of the start position throughout the test. Inadequate stabilization of the patient may allow movement of the limb and create variability in the measures of strength (Brown & Weir, 2001).

Adequate stabilization of the trunk also provides a stable base from which lower limb muscles can generate force (Hart, Stobbe, Till, & Plummer, 1984; Krause, Schlagel, Stember, Zoetewey, & Hollman, 2007; Stumbo et al., 2001). Various methods of stabilization have been employed in studies that have investigated measures of strength in the hip, with the use of seatbelts or a requirement for the participant to self-stabilize by holding on to the test surface being the most common. One study (Thorborg, Petersen, Magnusson, & Hölmich, 2010) that examined reliability of measures of hip strength with a HHD in asymptomatic participants, reported improved reliability when testing hip abduction and adduction strength in the supine position compared to the side-lying position. These authors suggested that the reduced variation found in this position was due to better stabilization of the patient. No belts were used to provide additional stabilization in this study, as the researchers wanted the measurement procedure to be “easy to learn, administer and implement in the clinical setting” (Thorborg et al., 2010, p. 497). Instead, participants were required to stabilize themselves by holding the plinth. Their results (ICC's between 0.74 and 0.98) indicate that this method has substantial reliability (Landis & Koch, 1977).

Of the studies that have examined strength of hip muscles using a HHD in people with hip pain, only two reported using stabilisation. Pua et al. (2008) used seat belts to stabilise the pelvis and opposite leg whilst Arnold et al. (2010) required participants to hold the edge of the plinth. The position of the participants also varied across these studies such that there was no one, consistent test position for most muscle groups e.g. Pua et al. and Arnold et al. tested hip flexors in sitting with the hip at 90° flexion whereas Sherrington and Lord (2005) tested this muscle group with the participant lying supine and the hip in neutral.

### **4.3.3 Make versus break**

HHD's allow measurement of the peak force generated by a muscle group. Two types of testing have been commonly employed in studies that have measured strength i.e. 'make' tests and 'break' tests. In a 'make' test, the examiner places the individual's limb in the desired position of testing, holds the HHD stationary and the person being tested is then required to exert the maximum force that they can generate against the firmly held HHD. With break tests, the examiner applies increasing force to the limb until the individual can no longer sustain the start position. Whilst there is a high correlation between the results of these two types of tests, 'break' testing has been demonstrated to show greater strength values than 'make' testing (Schmidt, Iverson, Brown, & Thompson, 2013; Stratford & Balsor, 1994). The high forces required to overcome an individual's ability to hold a given position during 'break' tests, may be unattainable in some circumstances (e.g. small female examiner versus strong male athlete) and create a risk of injury to the participant in others (Bohannon, 2012; Reiman & Thorborg, 2014).

Comparison of the reliability of 'make' versus 'break' techniques have been investigated in the hip (Schmidt et al., 2013). These authors measured hip abduction in 39 healthy subjects (aged 21 to 70) and reported that both methods were highly reliable. These results are consistent with the findings of other studies that have investigated this in the elbow (Bohannon, 1988; Stratford & Balsor, 1994). The three studies (Arnold et al., 2010; Pua et al., 2008; Sherrington & Lord, 2005) that investigated reliability of strength measures in patients with hip pain identified by the current search all used 'make' tests.

#### **4.3.4 Examiner strength**

There is some evidence that examiner strength influences the inter-examiner reliability of measures of strength in the hip. One study (Thorborg, Bandholm, Schick, Jensen, & Holmich, 2013) demonstrated a systematic difference in measurements of hip strength made by one female and one male examiner, with all measurements made by the female being lower than those of the male ( $p < 0.05$ ). Another study (Krause et al., 2014) reported similar findings with strength values being lower for the weaker examiner. Despite the difference in absolute values, these authors reported excellent intra-examiner reliability (ICC's between 0.82 and 0.97) and inter-examiner reliability (ICC's 0.81 to 0.98). Krause and colleagues argue that the excellent reliability, despite differences in strength values, reflects the mathematical analysis inherent with ICC's (the ratio of between-subject variance minus within-subject variance divided by between-subject variance). They conclude that this finding demonstrates that although such measurements are reliable, they are not necessarily valid. These authors recommended the use of long-lever techniques to minimize the influence of differences in examiner strength. Similarly, external fixation of the HHD has been suggested as a means to resolve this issue (Bohannon, 2012; Thorborg et al., 2010).

#### **4.3.5 Duration of contraction and rest intervals between repetitions**

Sufficient time needs to be given to allow for the development of peak force during isometric strength testing. It appears that there is not any experimental evidence to justify a specific time frame. However, the American Society of Exercise Physiologists (ASEP) guidelines (Brown & Weir, 2001) recommend that a contraction period with a one-second-transition period from rest to maximal force and a four to five second plateau is appropriate.

Similarly, sufficient time needs to be given for adequate recovery of muscle from the metabolic consequences of high-intensity, short duration muscle contractions.

Inadequate rest time has been demonstrated to impair subsequent performance (Spriet, Lindinger, McKelvie, Heigenhauser, & Jones, 1989). No definitive recovery period has been determined experimentally however the consensus of opinion based on the current evidence suggests that one minute is sufficient (Brown & Weir, 2001; Weir, Wagner, & Housh, 1994). Despite this evidence, it is clear that numerous studies that have examined strength of hip muscles have had recovery periods less than one minute (Arnold et al., 2010; Arokoski et al., 2002; Bandinelli et al., 1999; Bohannon, 1986; Harris-Hayes et al., 2014; Katoh & Yamasaki, 2009; Kelln, McKeon, Gontkof, &

Hertel, 2008; Krause et al., 2014; Morris, Dodd, & Morris, 2008; Phillips et al., 2000; Stockton et al., 2011; Thorborg et al., 2013). Only two of the three studies (Arnold et al., 2010; Pua et al., 2008; Sherrington & Lord, 2005) that investigated reliability of measures of hip strength in people with hip pain included in the Dobson et al. (2012) review, reported the duration of contraction and rest period. Arnold et al. performed a 5 second contraction with a 30 second rest whereas Pua et al. performed a 3-5 second contraction with a 60 second rest period.

#### **4.3.6 Number of repetitions**

Whilst the ASEP guidelines suggest that three repetitions are “sufficient to elicit a maximal value” during maximum isometric contractions, the findings from various studies demonstrate that one trial may be all that is necessary (Bohannon & Saunders, 1990; Brown & Weir, 2001; Coldham, Lewis, & Lee, 2006; Rasch et al., 2005).

Bohannon and Saunders investigated peak forces generated isometrically by elbow flexors. They compared the value obtained from the first maximum contraction to the ‘maximal’ value and to the mean value of three repetitions. These authors reported that all three methods demonstrated excellent reliability (ICC’s 0.97 to 0.98). However, a statistically significant difference ( $p = <0.001$ ) in values existed when the highest value was compared to the mean of the three repetitions. They concluded that a single repetition is likely to be sufficient and that it might be preferable when the population of interest is pathological so that the risk of aggravation of the underlying condition is minimised. These results are supported by those of Coldham et al. who reported similar findings with their investigation of maximum grip strength. Similarly, Rasch and colleagues reported no significant differences ( $p = <0.05$ ) in force values between the first and second maximum isometric contraction of hip muscles (flexors, extensors, abductors and adductors) with coefficients of variation (expressed as a percentage of the overall mean) between 3% and 6%.

Of the studies identified in the current search, Pua et al. (2008) used the mean value from 2 repetitions, Sherrington and Lord (2005) used the highest value from 2 repetitions and the third study (Arnold et al., 2010) appears to have used the average peak force from 3 repetitions of maximum voluntary contractions. Similar variety is seen amongst a more broad range of studies investigating reliability of measures of strength in normal hips and/or using instrumentation other than a HHD (Arokoski et al., 2002; Bloom & Cornbleet, 2014; Bohannon et al., 2008; Fulcher, Hanna, & Elley, 2010; Harris-Hayes et al., 2014; Herbert et al., 2011; Katoh & Yamasaki, 2009; Kelln et

al., 2008; Malliaras et al., 2009; Phillips et al., 2000; Thorborg et al., 2013; Thorborg et al., 2010; Wang, Olson, & Protas, 2002). It seems that there is not any experimental evidence that supports one particular method over another (Brown & Weir, 2001).

For measures of ROM there is conflicting evidence in respect to the effect of performing repeat measurements. Some studies have demonstrated that one measurement is as reliable as the mean of repeated measurements, whereas others have demonstrated improved reliability with repeat measures (Gajdosik & Bohannon, 1987). Two high quality studies (Holm et al., 2000; Pua et al., 2008) investigating the reliability of measures of ROM in people with hip pain were included in the systematic review by Dobson et al. (2012). Pua and colleagues reported that they used the mean of two repetitions whilst Holm et al. did not provide this detail. Considering other studies that have examined reliability of ROM measures in people with normal hips, there does not appear to be any clear preference or consensus with Arokoski, Haara, Helminen, and Arokoski (2004) performing a single measure, Malliaras et al. (2009) using the mean of two measures and two other studies (Nussbaumer et al., 2010; Prather et al., 2010) using the mean of three.

#### **4.3.7 Determining end of range**

When testing ROM passively, the limit of motion is determined by the amount of force applied to the limb. Some authors have recommended that this force should be standardised by using a force dynamometer to ensure that a consistent level of force is always applied (Gajdosik & Bohannon, 1987). A drawback with this method when investigating people with painful joints is that a pre-determined force may cause pain. This may have adverse effect on their condition and/or on the reliability of the measurement (Bierma-Zeinstra et al., 1998; Gajdosik & Bohannon, 1987; Pua et al., 2008; Steultjens, Dekker, Van Baar, Oostendorp, & Bijlsma, 2000).

Of the studies that have examined ROM in people with hip pain, only Pua et al. (2008) reported how they determined end ROM. These authors applied pressure to the point that a “firm or stiff end feel was felt”, unless pain restricted motion prior to this point. Of the studies that have examined reliability of range of movement in the normal hip, only Malliaras et al. (2009) reported this detail. These authors applied “gentle overpressure” to determine end ROM.

### 4.3.8 Summary

The evidence presented above suggests that measures of strength and ROM obtained by HHD's and inclinometers are sufficiently reliability to be used clinically. Whilst this may be the case, we considered that it was important to establish the reliability of measurements made by this researcher, using the device that will be used in the diagnostic accuracy study that is the focus of this thesis. The following decisions regarding the design and conduct of this study were determined based on the methodological considerations outlined in the previous text.

The use of external fixation was considered not to be pragmatic for the follow-up diagnostic accuracy study (Chapter 5 ) in which patients were examined before and after undergoing a magnetic resonance arthrogram (MRA). This study was set in a private radiology practice where restrictions in space and time meant that external fixation was not practical. Also, we wished to determine the reliability of strength testing performed in a manner that it can be easily performed in clinical practice (Thorborg et al., 2010). Consequently, participants were required to stabilise themselves by holding onto the plinth (in a standardised manner) during the performance of maximum voluntary isometric contractions. To be consistent with existing studies that have investigated the reliability of strength measures in patients with hip pain, 'make' tests were used rather than 'break' tests. Three maximum isometric contractions were performed, each with a 1-second ramp period followed by a 5-second hold. A two-minute rest period between each repetition was provided to allow for muscle recovery. The highest peak force measurement of the three repetitions was used for all subsequent analysis.

In respect to measures of ROM, a gravity dependent inclinometer was utilised. The participant was positioned in a standardised manner and carefully monitored so that any shift from the start position was observed and corrected. The use of a force dynamometer to apply a predetermined force to determine end ROM was considered but discarded for two key reasons. Firstly, participants in the follow-up diagnostic accuracy study were people with painful hip joint pathology. The possibility that the use of a predetermined force might be too provocative or detrimental for some participants was of concern. Secondly, the reproduction of the participant's pain is the outcome of interest with many diagnostic tests. Whilst too much load may be detrimental, a predetermined force may be *less* than that required to reproduce pain in some participants. Consequently, end range of movement was determined by the examiner feeling significant tissue resistance or by the patient stating that further motion would

be “unacceptably uncomfortable/painful”. The mean of three repetitions was used for subsequent analysis.

ROM of extension was not included in this study, primarily because previous research (Pua et al., 2008) has demonstrated that the absolute error (SEM%) associated with this measurement in people with hip pain is very high (74%). Additionally, we performed a small pilot study (n=3) to help develop and test data collection methods. This study revealed that it was very difficult for a single examiner, using an inclinometer, to measure ROM of extension passively and to obtain consistent values. Similarly, measurement of abduction ROM was difficult. Instead, we measured range of the ‘bent knee fall out’ (BKFO) test. This test is a measure of abduction and external rotation with the hip in approximately 45<sup>0</sup> of flexion. Restriction of motion in this direction is widely considered to be associated with hip joint pathology.

## **4.4 Methods and procedures of the current study**

### **4.4.1 Study design, participants and procedure**

This study was a test-retest study of the intra-examiner reliability of measures of strength and range of movement of the hip. Data for this study was collected simultaneously with that for the previous study (Chapter 3 ) using the same cohort of participants and by the same examiner. Thus, sample size calculation, inclusion criteria, baseline data collection and procedures were common to both studies and are therefore not repeated here (see page 58 for detail). In the current study, peak force and range of movement were measured rather than pain responses. Strength (kg) and range of motion (degrees) measurements were obtained using a previously validated hand held dynamometer (HHD) that incorporates both a force transducer and gravity dependent inclinometer (Industrial Research Ltd; Christchurch; NZ). The HHD was calibrated before data collection.

Standardised versions of all tests were performed bilaterally on each participant, with the asymptomatic hip tested before the symptomatic hip (for detail regarding performance of each test see Appendix 7). Strength measurements were made on the day of initial assessment (‘Session One’) and again several days later (average 4 days, range 2-7) in ‘Session Three’. Strength measures were not repeated during ‘Session Two’, which was performed 60 minutes after the initial session. Test order was standardised across both sessions for each participant. For each strength test, the participant was instructed how to perform the test and then a sub-maximal practice test

was performed. Next a 'practice' maximum voluntary contraction (MVC) was performed (the force produced was not measured). Finally, three repetitions of a MVC were performed with a 120 second rest between each repetition. During the measurement the examiner used a standardised instruction: "Go, push hard, push, push, push" to encourage the participant to produce a MVC and to make sure the contraction was maintained for a five second period. The contractions were all isometric 'make' force, performed against the HHD supported by the examiner. The participants were instructed to stabilise themselves by holding on to the testing plinth during the performance of each test. Pilot testing demonstrated that the participant's position did not change during testing and that the examiner was not overpowered by any participant. The force transducer measured the peak force that the participant generated during the five-second hold. Torque was not calculated given that we were comparing measurements obtained between-sessions from individual participants, not across participants. Hence, the length of the lever arm was consistent.

To measure range of motion, the examiner first placed the HHD on the body part to be moved (in a standardised fashion) and 'zeroed' the device such that the initial start position was determined with reference to the vertical plane. Next, the body part was moved through to its end range where the final position was recorded. End range of movement was determined by the examiner feeling significant tissue resistance or by the participant stating that further motion would be "unacceptably uncomfortable/painful". The HHD automatically subtracted the initial starting position from the final position and displayed the actual range of motion. Three repetitions were performed with approximately 30 seconds between measurements. Three sessions were performed with 'Session One' on the day of initial assessment and 'Session Two' approximately 60 minutes after Session One. 'Session Three' was 2-7 days later (dependent on participant availability).

#### **4.4.2 Analysis**

Normality of distribution was assessed using the Kolmogorov-Smirnov test and variables with a significance of greater than 0.05 were classified as having a normal distribution. For measures of strength, between-session reliability for symptomatic hips was calculated using the highest peak force measurement of the three repetitions performed in Session One (day 1), was compared to the highest peak force measurement of the three repetitions performed in Session Three (2-6 days later). Relative reliability was assessed using single-measure ICC's ( $ICC_{2,1}$ ) (two way random

and absolute agreement), for variables with a normal distribution, via the Statistical Package for the Social Sciences software, version 19.0 (SPSS Inc, Chicago, USA). Variables that did not exhibit normal distribution were assessed for reproducibility using Lin's Concordance Correlation Coefficient (CCC) (Dunn, 1992; Lin, 1989).

For measures of range of movement, both within-session and between-session reliability for symptomatic hips were calculated. Within-session reliability for each test was calculated by comparing the mean of the three scores from Session One (day 1) to the mean of the three scores from Session Two (60 minutes after Session One). The mean score for Session One was compared to the mean score from Session Three to calculate between-session reliability. Relative reliability was assessed using average-measure ICC's (ICC<sub>2,3</sub>) (two way random and absolute) via SPSS. The classification system of Landis and Koch (1977) was used for interpreting the ICC values i.e. ICC's of 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect.

Ninety five per cent confidence intervals were constructed as a measure of precision for ICC values (Sim & Wright, 2005). Absolute reliability (the degree to which repeated measures vary for individuals) was expressed as standard error of measurement (SEM) and were calculated for each test using the formula:  $SEM = SD \times \sqrt{1-R}$ , where the standard deviation (SD) was obtained from the data from Session One and the reliability (R) was the obtained reliability value (ICC) (Wyrwich, 2004). Percent SEM were also calculated by dividing the SEM with the average of the test results from Session One and Session Three for peak force, and from Session One and Session Two for ROM (Thorborg et al., 2010). Minimal detectable change (MDC) was calculated as  $SEM \times 1.96 \times \sqrt{2}$  (Weir, 2005).

Two tailed, paired t-tests were used to determine if there were any significant differences between-sessions in peak force or range of movement for the symptomatic hip. Similarly, these tests were employed to determine if there were any significant differences in peak force or range of movement between symptomatic and asymptomatic hips utilizing data from Session One.

## 4.5 Results

Baseline and descriptive data for the participants included in this study were presented in Chapter Three (see section 3.4. for detail). The following results are specific to the measurements of strength and ROM.

### 4.5.1 Reliability of strength testing measurements

Peak isometric force was measured for each participant for each resisted test. Table 4.1 shows the mean values and standard deviation (SD) across participants for each test for each of the sessions where maximum voluntary contractions were performed (Session One and Three). Mean forces ranged from 14.5 kg for adduction to 32.8 kg for flexion. For each of the resisted tests, the mean peak force generated in 'Session Three' was higher than that generated in 'Session One'. The difference in values between these two sessions ranged from 0.6 to 1.7 kg. These differences were statistically significant ( $p \leq 0.05$ ) for three of the six tests (abduction, flexion and internal rotation). Resisted adduction and external rotation tests demonstrated normal distribution and were therefore assessed for reliability using ICC's. The other resisted tests were assessed using Lin's CCC. Table 4.1 also provides detail regarding concordance values for all of these tests. Values ranged from 0.70 (95% CI 0.36, 0.87) for adduction to 0.92 (95% CI 0.81, 0.97) for extension. All resisted tests demonstrated 'substantial' or 'almost perfect' reliability. Errors associated with these measures are detailed in Table 4.1. The standard error of the measurement (SEM) ranged from 1.5 to 3.0 kg. SEM percent was highest for adduction (16%). This was reflected in the minimal detectable change (MDC) for this muscle group. Whilst MDC percent for most resisted movements was around 30%, a much higher value was seen with adduction (44%).

Peak isometric force was also measured for the asymptomatic hips. This enabled comparison of strength values between symptomatic and asymptomatic hips. Table 4.2 shows the mean values and SD for data collected at Session One based on the 16 participants that had both an asymptomatic hip and symptomatic hip (i.e. two participants with bilateral hip pain were excluded from this analysis). The strength differences between the two hips were not statistically significant for any test. This was also the case for data collected at the follow up testing session (data not shown).

**Table 4.1 Between-session mean peak force measurements for symptomatic hip (n=18)**

Resisted Test	Mean Peak Force in Kg (SD) Session One	Mean Peak Force in Kg (SD) Session Three	Mean Difference <sup>1</sup> in Kg (SD)	p-value	CCC/ICC <sub>2,1</sub> (95% CI)	SEM in Kgs	SEM%	MDC in Kgs	MDC%
Abduction	15.4 (4.0)	16.8 (4.6)	1.4 (2.3)	0.02*	0.82 (0.61,0.92)	1.7	10.5	4.7	29
Adduction	14.5 (4.3)	15.4 (4.6)	0.9 (3.4)	0.29	0.70 (0.36,0.87)	2.4	16.0	6.6	44
Extension	19.3 (6.4)	19.9 (6.1)	0.6 (2.3)	0.33	0.92 (0.81,0.97)	1.8	9.2	4.9	25
Flexion	31.1 (6.2)	32.8 (4.5)	1.7 (3.3)	0.05*	0.77 (0.54,0.89)	3.0	9.4	8.3	26
Internal Rotation	19.4 (5.1)	21.1 (4.7)	1.7 (2.9)	0.02*	0.78 (0.52,0.91)	2.4	11.8	6.6	33
External Rotation	14.8 (4.2)	15.6 (4.2)	0.9 (2.0)	0.09	0.87 (0.67,0.95)	1.5	9.8	4.1	27

<sup>1</sup>Mean difference in force between sessions; \* Statistically significant difference between-sessions ( $p \leq 0.05$ ); Kg, kilograms; CI, confidence intervals; SD, Standard Deviation; CCC/ICC, concordance correlation coefficient/intraclass correlation coefficient; SEM, standard error of measurement; SEM%, standard error of measurement expressed as a percent of mean values of both Session One and Session Three; MDC, Minimal detectable change; MDC%, Minimal detectable change expressed as a percent of mean values of both Session One and Session Three

**Table 4.2 Peak force measurements for asymptomatic versus symptomatic hip<sup>1</sup> (n=16)**

Resisted Test	Asymptomatic Hip Mean Peak Force in Kg (SD)	Symptomatic Hip Mean Peak Force in Kg (SD)	Mean Difference <sup>2</sup> in Kg (SD)	% Difference <sup>3</sup>	p-value
Abduction	16.4 (5.0)	15.7 (3.8)	-0.73 (2.3)	4.4	0.22
Adduction	14.9 (4.0)	15.0 (4.1)	0.06 (2.8)	0.4	0.94
Extension	19.8 (6.9)	19.8 (6.3)	0.01 (2.2)	0.05	0.99
Flexion	32.3 (6.0)	31.9 (5.0)	-0.41 (2.7)	1.3	0.55
Internal Rotation	19.4 (5.1)	19.9 (4.8)	0.55 (2.7)	2.8	0.44
External Rotation	15.0 (3.4)	15.2 (4.0)	0.20 (2.0)	1.3	0.75

<sup>1</sup>Based on Session One data. SD, Standard Deviation; Kg, kilograms; <sup>2</sup>Mean difference between asymptomatic and symptomatic hip; <sup>3</sup>Mean difference expressed as a % of mean peak force of the asymptomatic hip.

#### 4.5.2 Reliability of ROM measurements

Table 4.3 shows the *within*-session mean values and SD across participants for each of the range of movement tests for the symptomatic hip. The mean differences between-sessions were not significant for any test. ROM data for flexion, internal and external rotation demonstrated normal distribution and were therefore assessed for reliability using ICC's. The bent knee fall out (BKFO) test was assessed using Lin's CCC. All tests demonstrated 'almost perfect' reliability. Error associated with these measurements was highest for internal rotation with a SEM percent of 6% and MDC percent of 16.5%.

Table 4.4 provides detail for the *between*-session testing of the symptomatic hip. The mean differences between-sessions were not significant for any of the tests. Concordance values ranged from 0.82 for flexion to 0.95 for the BKFO. All between-session ROM tests demonstrated either 'substantial' or 'almost-perfect' reliability. Absolute error was higher for all between-session measures than that seen for within-session measurements.

Range of movement was also measured for the asymptomatic hip. This enabled comparison of range between symptomatic and asymptomatic hips. Table 4.5 shows the mean values and SD for data collected at Session One based on the 16 participants that had both an asymptomatic hip and symptomatic hip (i.e. two participants with bilateral hip pain were excluded from this analysis). There was a statistically significant mean difference of 3.5 degrees for the BKFO test, with the symptomatic hip demonstrating less ROM than the asymptomatic hip. The mean differences between the two hips were not significant for any of the remaining tests.

**Table 4.3 Within-session ROM measurements in degrees for symptomatic hip (n=18)**

ROM Test	Mean ROM (SD) Session One	Mean ROM (SD) Session Two	Mean Difference <sup>1</sup> (SD) in degs	<i>p</i> -value	CCC/ICC <sub>2,3</sub> (95% CI)	SEM in degs	SEM%	MDC in degs	MDC%
BKFO	61 (9.1)	62 (9.8)	1.0 (3.1)	0.20	0.97 (0.92, 0.98)	1.6	2.6	4.4	7.1
Flexion	114 (11.0)	115 (11.8)	0.9 (3.9)	0.32	0.97 (0.92, 0.98)	1.9	1.6	5.2	4.5
Internal Rotation	36 (9.2)	34 (9.8)	1.6 (3.9)	0.10	0.95 (0.86, 0.98)	2.1	6.0	5.8	16.5
External Rotation	41 (7.2)	42 (8.5)	0.3 (3.2)	0.66	0.96 (0.89, 0.98)	1.5	3.6	4.1	9.8

ROM, Range of Movement; degs, degrees; CCC/ICC, concordance correlation coefficient/intraclass correlation coefficient; SD, Standard Deviation; BKFO, bent knee fall out;

<sup>1</sup>Mean difference in ROM between sessions; CI, confidence intervals; CCC/ICC, concordance correlation coefficient/intraclass correlation coefficient; SEM, standard error of measurement; SEM%, standard error of measurement expressed as a percent of mean values of Session 1 and Session 2; MDC, Minimal detectable change; MDC%, Minimal detectable change expressed as a percent of mean values of Session 1 and Session 2

**Table 4.4 Between-session ROM measurements in degrees for symptomatic hip (n=18)**

ROM Test	ROM (SD) Session Two	ROM (SD) Session Three	Mean Difference <sup>1</sup> (degs)	<i>p</i> -value	CCC/ICC <sub>2,3</sub> (95% CI)	SEM in degs	SEM%	MDC in degs	MDC%
BKFO	62 (9.8)	61 (7.9)	1.1 (3.8)	0.24	0.95 (0.87, 0.98)	2.2	3.5	6.0	9.7
Flexion	115 (11.8)	113 (13.1)	2.1 (9.8)	0.37	0.82 (0.51, 0.93)	5.0	4.3	13.8	12.0
Internal Rotation	34 (9.8)	36 (10.4)	2.1 (5.9)	0.15	0.90 (0.74, 0.96)	3.1	9.1	8.6	25.3
External Rotation	42 (8.5)	41 (7.6)	0.5 (3.9)	0.60	0.94 (0.84, 0.97)	2.1	5.0	5.8	13.8

ROM, Range of Movement; degs, degrees; CCC/ICC, concordance correlation coefficient/intraclass correlation coefficient; SD, Standard Deviation; BKFO, bent knee fall out;

<sup>1</sup>Mean difference in ROM between sessions; CI, confidence intervals; CCC/ICC, concordance correlation coefficient/intraclass correlation coefficient; SEM, standard error of measurement; SEM%, standard error of measurement expressed as a percent of mean values of Session 2 and Session 3; MDC, Minimal detectable change; MDC%, Minimal detectable change expressed as a percent of mean values of Session 2 and Session 3

**Table 4.5 ROM measurements in degrees for asymptomatic versus symptomatic hip<sup>1</sup> (n=16)**

ROM Test	Asymptomatic Hip Mean ROM (SD)	Symptomatic Hip Mean ROM (SD)	Mean Difference <sup>2</sup> in degs	<i>p</i> -value
BKFO	65 (6.8)	61 (9.0)	3.5	0.03*
Flexion	116 (11.0)	114 (11.2)	0.9	0.44
Internal Rotation	39 (6.2)	36 (9.5)	2.4	0.18
External Rotation	42 (5.2)	41 (7.7)	0.6	0.64

<sup>1</sup>Based on Session One Data; <sup>2</sup>Mean difference in ROM between hips; \* Statistically significant difference between sessions ( $p \leq 0.05$ ); ROM, Range of Movement; degs, degrees; SD, Standard Deviation; BKFO, bent knee fall out.

## 4.6 Discussion

The primary aim of this study was to investigate the intra-examiner reliability of measures of strength and range of movement in people with hip pain, using a HHD/inclinometer, to determine if such measures could be employed with confidence in the diagnostic accuracy study (Chapter 5 page 118). This study has provided evidence that supports and extends previous research in regard to the both relative and absolute reliability of such measures. It has also provided new and important information in respect to the degree of change needed to recognise a ‘true’ difference in strength or ROM. It is the first study to investigate reliability of the measurement of adductor strength and of the ROM of the bent knee fall out in people with hip pain. The following discussion is structured around each of the key measures investigated in the study.

### 4.6.1 Strength measurements

#### Reliability

Our results demonstrate that between-session intra-examiner measurements of peak isometric force measured with a HHD in people with a painful hip joint have excellent levels of reliability (ICC values ranged from 0.70 to 0.92). These results are consistent with previous good quality studies (Arnold et al., 2010; Pua et al., 2008; Sherrington & Lord, 2005) that have investigated these measures in a similar cohort. Whilst there is not complete homogeneity between each of these studies and our own in respect to included pathologies, methodology and statistical analysis, comparison of their results with ours can be made with caution (Haas, 1991).

ICC values reported by Pua et al. (2008) ranged from 0.84 to 0.98, slightly higher for each test than our findings. These authors calculated reliability by using the mean of two peak torque measurements whereas we used the *single* highest peak force generated by each participant. Mean values ‘average out’ errors associated with repeat measures leading to higher reliability values (Hopkins, 2000). This may have contributed to the slightly higher reliability in their study. Pua et al. considered that the primary reason for the high reliability they observed was that they used seat belts to stabilise their participants during testing. However, our findings indicate that reliable test-retest measures of hip strength can be made without the need for such stabilisation devices. This is supported by the findings of Arnold et al. who employed the same stabilisation strategy as us. These authors reported ICC values ranging between 0.90 and 0.98 for

measures of hip abduction, flexion and extension strength. In contrast to our study, but in a similar fashion to Pua et al., Arnold and colleagues used the mean of two measurements of peak force to calculate the reliability. ICC values reported by Sherrington and Lord are very similar to our own i.e. 0.80 (versus our 0.77) for flexion strength and 0.86 (versus 0.82) for abduction strength. These authors used the single, highest peak force measurement to calculate reliability in the same manner as in the current study.

We reported estimates of absolute reliability using the standard error of measurement (SEM) as did both Pua et al. and Arnold et al. The SEM is expressed in the same unit that the actual measurements were made, therefore making it easy to interpret the size of the error. However, when comparing findings between studies, this is not useful unless the same unit of measurement has been used. In both the current study and that of Arnold et al., strength was measured in kilograms whereas Pua and colleagues reported torque (in Newton metres). To enable comparison between these studies and our own, we calculated the percent SEM from the data reported by these other researchers.

Abduction, flexion and extension strength were investigated in all 3 studies. Percent SEM for abduction in our study was 10.5%, very similar to both Arnold et al. (10.4%) and Pua et al. (13.6%). Similarly, the results for flexion across these studies were comparable (9.4%, 13.1% and 10% respectively). With respect to measures of extensor strength, the error of 17.3% in the study by Arnold and colleagues was much larger than the 8.6% reported by Pua et al. and the 9.2% in our study. SEM% for measures of internal and external rotation strength in our study were 11.8% and 9.8% respectively, very similar to those of Pua et al who reported 8.3% and 7.5%. Apart from the large error associated with extensor strength measures reported by Arnold et al., the SEM% across all three studies and all muscle groups are close to 10%. This consistency, despite the various differences in the characteristics of included participants and methods employed across studies, provides some confidence in the legitimacy of these findings. Hence, changes in measures of strength of 10% or less may well be a result of measurement error, suggesting that 'real' strength changes cannot be confidently appreciated unless there is more than a 10% difference in test values.

Interestingly, the error associated with measurements of adduction strength in our study was higher than for all other strength tests, with a SEM% of 16%. Although the point estimate (ICC=0.70) of relative reliability for this measurement falls within the

‘substantial’ range, the lower confidence interval was only 0.36. This high SEM%, combined with the wide confidence intervals of the ICC findings suggests that changes in isometric strength measures of this muscle group should be interpreted with caution. Unfortunately, we cannot easily compare this finding to previous research, as it appears that no other study has investigated the reliability of measures of adductor strength in people with hip pain. Thorborg et al. (2011a) performed a pilot study (n = 10) to determine intra-examiner reliability of *eccentric* hip adduction and adduction strength in people *without* any history of hip pain. These researchers reported a SEM% of just 6.3% for adduction (and 5.1% for abduction), a value much lower than our own. Thorborg et al. (2010) observed low SEM% values (ranging from 3 to 8%) in another small study (n = 9) of healthy participants in which they performed *isometric* strength testing. The differences in the magnitude of the error between these studies and our own may be related to the presence of painful pathology in our subjects.

Our results, along with these previous studies, indicate that repeat measures of strength using a HHD have excellent relative and absolute reliability for hip muscle groups other than the adductors. This appears to be the case despite various methods of measurement including the actual device used, the manner of stabilisation and the position of the patient. A priori, we had considered that such measures in people with pathological hips might be unreliable given the presence of pain. Although many participants in the current study reported pain during strength testing (see Chapter 3, Table 3.8 for detail), the presence of pain was not enough of a factor to make these tests unreliable for the majority of tests.

### **Minimal detectable change**

Whilst the SEM is an estimate of the amount of error associated with a measurement, the minimal detectable change (MDC) is a statistical estimate of the smallest amount of change in measurement score (value) necessary before one can be confident that a true change in patient status has occurred, as opposed to a change in scores resulting purely from error associated with the measurement (Schmitt & Di Fabio, 2004). Similar to the SEM, the MDC is expressed in the same units as the outcome measure. In the current study, MDC values were calculated using a 95% confidence level and ranged from 4.1 kg for external rotation strength to 8.3 kg for flexion. To enable comparison across the different strength tests investigated in the current study, and with findings of other studies that have investigated measures of hip strength in people with hip pain, we have

also provided MDC values as a percent of the mean values obtained across the two test sessions.

Of the three studies (Arnold et al., 2010; Pua et al., 2008; Sherrington & Lord, 2005) that have examined the reliability of strength measures in people with hip pain, only Pua et al. reported MDC values. However, these authors calculated the MDC using 90% rather than 95% confidence intervals. Although neither Arnold et al. nor Sherrington and Lord reported MDC, they provided sufficient data to enable calculation of MDC and MDC%. Similarly, MDC% based on 95% confidence intervals could be calculated from the data provided by Pua et al. For abduction, the MDC% values across these studies ranged from 29% (Arnold et al.) to 51% (Sherrington and Lord). In our study this value was 29%. For flexion, we reported 26%, compared to a range from 28% (Pua et al.) to 82% (Sherrington and Lord). The MDC% for extension in our study was 25%, similar to Pua et al (24%) and much smaller than the 48% reported by Arnold et al. Only Pua and colleagues investigated internal and external rotation, reporting 23% and 21% respectively compared to our 33% and 27%. Our findings generally sit toward the lower end of the range of values seen across these studies. The values demonstrated in the study by Sherrington and Lord are much higher than those from the other studies, most likely as a result of the relatively old age of the patients (mean age  $79 \pm 10$  years) and the type of pathology involved (post hip fracture). Another factor likely to have increased the MDC% values in their study was that data was collected anywhere between 1 and 57 weeks post-fracture. Hence, participants were at various stages of recovery and likely to have a wide range of strength values.

Whilst these other studies provide some context for our own results, it is difficult to make direct comparisons due to the numerous differences in both methods and participant characteristics as previously discussed. Our findings, along with those of previous researchers, suggest that the MDC% required to determine that a true and meaningful change in muscle strength has occurred as a result of a strengthening program varies according to the muscle group being tested, the manner of testing and the patient characteristics. We demonstrated that a MDC% close to 30% is required for all hip muscle groups except adduction (44%). Changes in strength below these values may represent error associated with the measurement. These values provide benchmarks for clinicians working in the primary health care setting that will allow them to confidently interpret changes in patient strength status.

### **Side-to-side strength**

In our study, we did not find any statistically significant differences in muscle strength between the symptomatic and asymptomatic hips. This was surprising considering that the average length of time that symptoms had been present was 25 months and that over a third of participants had significantly modified or stopped their normal daily activities because of their hip pain. One would expect a degree of weakness of the muscles around the painful joint secondary to muscle atrophy or arthrogenic muscle inhibition associated with pain, swelling or inflammation (Rice & McNair, 2010).

Our results for abduction and adduction support those of Arokoski et al. (2002), who reported that there were not any differences between sides for hip abductors or adductors in the patients with OA included in their study. However, these researchers did see a 22% reduction in strength of the flexors and extensors in the symptomatic hip that was statistically significant. Our results also contrast with those of Rasch et al. (2007) who investigated patients on a waiting list for hip arthroplasty due to end-stage OA (average age 69 years). These authors measured strength with a strain-gauge dynamometer and reported a mean decrease of 19% of the symptomatic hip compared to the asymptomatic hip and statistically significant differences for flexion, extension, abduction and adduction. Twenty of the 22 patients investigated by these authors were followed up over a two-year period post hip arthroplasty. Rasch et al. (2010) reported on this cohort, stating that statistically significant difference remained in all muscles at 6 months but only in the abductors at 24 months.

The most likely reason for our findings contrasting with those of these previous studies are differences in the severity of the condition causing the participants hip pain. The patients in the studies by Rasch and colleagues had a mean age of  $67 \pm 7$  years and had undergone total hip arthroplasty, suggesting that they had severe end-stage OA. Forty percent of patients included in the Arokoski et al. study had radiological evidence of 'moderate' OA and symptoms that had been present for a mean of  $6.4 \pm 5.2$  years. The mean age of their participants was  $56 \pm 4.9$  years. The participants in our study were relatively young (mean age  $29.5 \pm 8.51$ ) and their symptoms had only been present for an average of  $2.1 \pm 1.7$  years. Our participants reflect patients with hip pain in the primary health care environment and were likely have less severe pathology. Our results suggest that people with hip pain do not develop significant weakness of the muscles of painful hip in the early stages of the 'disease' process (or post injury).

An alternative explanation of our findings may be that a similar degree of weakness was present in both hips such that side-to-side differences were not identified. It is not unrealistic to expect that people with hip pain could develop bilateral weakness as a consequence of restricting their activities of daily living and recreation as a result of any pain associated with these activities. This suggestion is supported by the findings of Harris-Hayes et al. (2014) who demonstrated weakness in the asymptomatic hip of people with unilateral hip pain when compared to control subjects. If this second interpretation is correct, then the utility of comparing side-to-side strength in expectation that unilateral weakness suggests a pathological hip joint should be questioned. This contradicts the recommendation of Kemp et al. (2013) who concluded, on the basis of their study of people with normal hips, that “that the use of the unaffected limb as a comparator when examining patients with unilateral lower-limb injury would be considered reasonable”.

#### 4.6.2 ROM Measurements

##### Reliability

Our results demonstrate that within and between-session measurements of range of movement measured with an inclinometer in people with a painful hip joint have almost perfect levels of intra-therapist reliability (within-session CCC/ICC's 0.95 to 0.97 and between-session CCC/ICC's 0.82 to 0.95). Our results are consistent with those of Pua et al. (2008), the only other study that has examined the *intra*-examiner, between-session reliability of measures of ROM in people with hip pain that we are aware of. Pua et al. examined reliability of measures obtained with an inclinometer and reported ICC's between 0.86 (95% CIs 0.67, 0.94) and 0.97 (95% CIs 0.93, 0.99).

Absolute error in the current study, expressed as the percent of SEM (SEM%), ranged between 1.6% (for flexion ROM) and 6.0% (for internal rotation). By comparison, we calculated SEM% from the data provided by Pua et al. In their study, SEM % ranged from 1.8% (for abduction) to 74% for extension. The very large SEM% for extension reflects the relatively large SEM compared to the very small ROM for extension ( $6.2^{\circ} \pm 10.7^{\circ}$ ). For the three movements investigated in both the current study and that of Pua et al. (flexion, internal and external rotation), our results demonstrated smaller percent error (1.6%, 6% and 3.6% respectively) than Pua and colleagues (2.9%, 11.1%, and 7.2%). The smaller error associated with our study may reflect a difference in the method of calculation of mean scores given that we used a mean of three measures whereas these authors used a mean of two. The error associated with these measures is

taken into consideration in the calculation of the minimal detectable change (MDC). We reported MDC values ranging from  $4.1^{\circ}$  for flexion to  $6^{\circ}$  for internal rotation. Expressed as a percentage of the raw values obtained in testing (i.e. MDC%) our results suggest that a true change in ROM cannot be assumed to have occurred unless there is more than a 7% increase (or decrease) in ROM for the BKFO test. Similarly, increases/decreases of greater than 5%, 16% and 9.8% are required to be sure that a true change has occurred for flexion, internal and external rotation respectively.

### **Range of movement**

We can compare our results for the actual range of movement measured in our study with those of both Pua et al. and Klassbo et al. (2003). The mean ROM for flexion in the current study was  $114^{\circ} \pm 11^{\circ}$ , comparable to the  $117^{\circ} \pm 13.9^{\circ}$  reported by Pua et al. and  $110^{\circ} \pm 17.3^{\circ}$  reported by Klassbo et al. For external rotation, the mean range in our study was  $41^{\circ} \pm 7.2^{\circ}$ , very similar to Pua and colleagues ( $42.8^{\circ} \pm 12^{\circ}$ ) but very different to Klassbo ( $21^{\circ} \pm 12.3^{\circ}$ ). Our results contrast to both of these other studies for internal rotation where we measured  $36^{\circ} \pm 9.2^{\circ}$  compared to the  $30.6^{\circ} \pm 9.4^{\circ}$  in the Pua study and the  $22^{\circ} \pm 13.8^{\circ}$  reported by Klassbo et al. This difference in internal rotation range is not unexpected given that both of these other studies included older patients, all of whom had radiological evidence of hip OA. Whilst our study had some older patients who may have had OA, we purposely included a broader range of patients with hip pain to reflect the cross section of patients who present in the primary health care environment. Thus, we also included younger patients with suspected labral pathology and/or femoroacetabular impingement.

To provide further context to our results, comparison of ROM in people with 'normal' hips is useful. Klassbo et al. (2003) examined range in a large cohort (n=177 hips) of people without pain. In this group flexion was  $122^{\circ} \pm 11.9^{\circ}$  (range 104-142), and internal and external rotation were  $34^{\circ} \pm 10^{\circ}$  (range 4-51) and  $28^{\circ} \pm 9.9^{\circ}$  (range 6-36) respectively. Roaas and Andersson (1982) reported very similar findings in a study that measured ROM with a goniometer in 210 hips. In this study, flexion was  $120^{\circ} \pm 8.3^{\circ}$  (range 90-150), and internal and external rotation were  $32.6^{\circ} \pm 8.2^{\circ}$  (range 20-50) and  $33.6^{\circ} \pm 6.8^{\circ}$  (range 10-55) respectively. There appears to be some agreement between these studies in terms of mean ROM for both flexion and internal rotation. However, the range of values demonstrates that range of motion in asymptomatic hips varies widely. The ranges that we have reported are not dissimilar to those reported in these normative

studies, except for external rotation where we saw a much larger mean ( $41^{\circ}$ ) and a narrower range. There is no obvious reason for why this was the case in our study.

We observed a mean range of movement for the BKFO test of  $61^{\circ} \pm 9.1^{\circ}$  (range 43-74). Although this test is often employed during the clinical examination of people with a painful hip, we are unaware of any other studies that have reported the ROM for this test. Our findings provide a benchmark for comparison in future studies and clinical practice.

Our comparison of mean values of ROM between symptomatic and asymptomatic hips is novel. We observed a statistically significant reduction in range for the BKFO test on the symptomatic side. The mean difference was 3.5 degrees, greater than the SEM ( $1.6^{\circ}$ ) associated with this measurement but less than the calculated minimal detectable change (MDC) of 4.4 degrees. We are unaware of any other studies that have reported side-to-side differences in people with unilateral hip pain.

#### **4.7 Limitations**

This study was a pragmatic study designed to determine the intra-examiner reliability of measures of strength and ROM in a manner that would allow collection of such data to be utilised in the follow-up diagnostic accuracy study. Hence, all measurements were made by a single examiner (the researcher) who was not be blinded to the values obtained with each test. To minimise any potential bias associated with these measures, the HHD/inclinometer was positioned in such a manner that the values could not be seen during the actual test manoeuvre. The device we used ‘captures’ and retains the highest peak force and/or largest range of motion achieved during the test manoeuvre. This allows the examiner to remove the device from the patient, and therefore to read and record the value after the test has been completed. Similarly, it was not possible to blind the examiner to the pathological status of the hip in this study. Given this knowledge, it may be that the examiner unintentionally modified the degree of force applied to determine end range of movement during ROM testing. However, this is unlikely given that there were not any statistically significant differences between sides in any of the peak force or ROM values, other than for ROM of the BKFO test.

The participants in our study were relatively young (mean age 29.5 years) volunteers with unilateral hip joint pain that had been present for an average of two years. Whilst our cohort did include 6 participants diagnosed with osteoarthritis, the majority were

diagnosed with FAI/labral tears. Our findings may not be applicable to older patients with degenerative conditions that have been present for many years. However, they are representative of the mix of patients who are likely to present at primary health care clinics and of the cohort that will be included in the following diagnostic accuracy study.

This study was powered for the primary purpose of the chapter i.e. to determine the reliability of measures of strength and ROM in patients with hip joint pain. The sample size is a limitation in respect to our findings regarding differences in strength and ROM between sides.

#### **4.8 Conclusions and implications**

Despite the presence of pain and pathology, this study has demonstrated ‘almost perfect’ levels of intra-examiner reliability for measures of ROM and ‘substantial’ to ‘almost perfect’ levels for measures of peak isometric force. The methods we employed to obtain these measurements are pragmatic and easily transferable to the clinical environment. Provided the error associated with these measurements is considered, we can recommend that changes in strength and ROM can be used clinically to determine changes in patient status as a result of an intervention.

This study also provided evidence of this examiners ability to make reliable measures of strength and ROM, justifying the inclusion of such measures in the follow-up diagnostic accuracy studies. However, the study highlighted that the time necessary to measure peak force in a reliable manner is considerable. With the ‘warm up’ and familiarisation period, three repetitions of maximum voluntary contractions (with a two-minute recovery period between repetitions) and six muscle groups to test, this process added at least 40 minutes to time taken for data collection.

Initially, we had intended to investigate the diagnostic accuracy of changes in hip strength for the identification of intra-articular pathology in the diagnostic accuracy study reported in the following chapter. The diagnostic accuracy study was performed in a private radiological practice on patients referred for a fluoroscopy-guided injection of anaesthetic and a magnetic resonance arthrogram. Data collection had to be performed on site, immediately before these medical procedures and repeated immediately after their conclusion. Co-ordination between the researcher, the patient, administrative staff, nursing and medical professionals in this environment was crucial to successful data collection and on-going recruitment. Consultation with the relevant

personnel in this practice determined that the time available for data collection was restricted. Given these concerns, measures of peak force were *not* included in the follow-up study.

## **Chapter 5 The Diagnostic Accuracy of Findings from the Clinical Examination of the Painful Hip**

This chapter relates specifically to Question 4 of this thesis:

*How accurately do individual findings obtained from the clinical examination of the hip predict a positive response to an intra-articular injection of anaesthetic into the hip joint?*

### **5.1 Introduction and Background**

An essential component of the diagnostic process is the collection and interpretation of information from the history and physical examination of the patient (Feddock, 2007; Peterson et al., 1992; Woolf, 2003). Such information enables the examiner to consider the combination of symptoms and signs observed in the patient that they are examining and to compare them to those associated with established diagnoses with which they are familiar. These initial diagnoses are typically ranked in terms of most to least likely and should direct the subsequent examination. Each piece of new information (test result) has the potential to cause the examiner to modify their initial hypotheses. A given test result may provide evidence that supports a particular diagnosis and will therefore increase the examiners estimate of the likelihood of that diagnosis being present. Alternatively, a test finding may indicate that a preferred hypothesis is now less likely. Hence, it is essential that information collected during the patient examination is valid and that it is interpreted correctly. Because few tests are 100% accurate, knowledge of the actual level of accuracy of a particular test allows the examiner to consider the diagnostic value of the information gained from that test. A number of systematic reviews that have investigated the diagnostic accuracy of the clinical examination of the hip have stated that there is currently insufficient evidence to make any valid conclusions regarding the value of such information (Burgess et al., 2011; Rahman et al., 2013; Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012).

This issue is important. There is evidence to suggest that an incorrect or delayed diagnosis may allow time for further deterioration of the patient's condition (Ganz et al., 2008; McCarthy et al., 2001). Already, the costs of treatment of hip joint injuries to the Accident Rehabilitation, Compensation and Insurance Corporation (ACC) in New Zealand ACC are substantial, currently in the vicinity of eighty million dollars a year (Accident Compensation Corporation, 2015). A well-designed study that provides

definitive estimates of the accuracy of the information obtained from the clinical examination of the hip will allow clinicians to confidently interpret the findings of such information. This will enable them to make a more timely and accurate diagnosis and to initiate appropriate management expeditiously.

Therefore, this chapter focuses upon the diagnostic accuracy of such information. First, a literature review that considers previous relevant research is presented. Next, key methodological factors important to the design and conduct of a diagnostic accuracy study will be considered. Finally, the methods and results of the diagnostic accuracy study undertaken as a part of this thesis will be reported and discussed.

## **5.2 Literature review**

The general aim of this review was to identify any previous relevant research that would inform the design and conduct of the study reported in this chapter. The more specific aim was to identify and consider studies that had investigated the diagnostic accuracy of physical tests for intra-articular pathology of the hip. The initial search was performed in July 2011, prior to the commencement of data collection for the current study, utilising the search strategy detailed in Chapter 2 . Because this search was conducted to inform the proposed study, the inclusion criteria were kept relatively broad i.e.

- Any study that investigated the diagnostic accuracy of clinical tests of the hip against an appropriate reference standard.
- Any study that investigated associations between findings of the clinical examination (history or tests) of the hip and an appropriate reference standard.
- Narrative reviews or expert commentaries considering the diagnostic accuracy of the clinical examination of the hip
- Any study that investigated the diagnostic accuracy of imaging techniques for the diagnosis of hip pathology

This search identified two relevant systematic reviews (Burgess et al., 2011; Leibold et al., 2008) that between them included 27 relevant publications. Fourteen publications that were not included in either of the two systematic reviews were also identified with the initial search. Additional publications were subsequently included in this review as a result of being identified by RSS feeds (linked to the original search), through a follow-

up search performed on the 24<sup>th</sup> March 2015 and by hand searching of references from retrieved papers.

Table 5.1 provides detail with respect to the search terms used and results for each database for the follow-up search (24<sup>th</sup> March 2015). This search identified 937 citations. All titles and abstracts were screened to determine relevance. Full-text versions of any publication for which relevancy could not be determined from the abstract were retrieved.

**Table 5.1 Overview of search terms and results per database**

Search Terms	SCOPUS	AMED	Medline via EBSCO	Medline via PUBMED	SPORT Discus via EBSCO	CINAHL via EBSCO
1 "hip joint" OR "hip pain" OR "femoroacetabular impingement" OR "FAI" OR labr* OR (osteoarthritis* N5 hip*) OR (OA N5 hip) OR (arthrit* N5 hip*) OR "ligamentum teres"	56,440	1517	52,656	44,599	7,400	9,625
2 accura* OR sensitivity OR specificity OR validity OR "likelihood ratio"	4,199,223	12,137	1,811,148	1,916,761	39,832	188,240
3 "Physical examination" OR "Orthopaedic Tests" OR "Pain provocation tests" OR "Objective examination" OR "Special Tests" OR "Impingement Test" OR FABER OR "Range of Movement"	179,322	1,384	73870	74,068	2,893	25,925
1 and 2 and 3	280	112 <sup>1</sup>	193	203	43	106
Total Number of Titles Identified = 937						

<sup>1</sup> 1 and 2 only as 1 and 2 and 3 narrowed to just 9

Four additional systematic reviews were identified in this search (Rahman et al., 2013; Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012). These systematic reviews included 13 publications that were not identified by the initial search (July 2011). With respect to diagnostic accuracy studies for clinical tests of the hip, a total of 18 original studies were identified and all but three (Chong, Don, Kao, Wong, & Mitra, 2013; O'Donnell et al., 2014a; Ochiai, Adib, & Donovan, 2011) were included in at least one of the systematic reviews. Two of these studies (O'Donnell et al. and Chong et al.) were published after these reviews. The other study (Ochiai et al., 2011) has only been published as an abstract of a conference presentation. Considering the quality of

these systematic reviews (see following section entitled ‘Identified Systematic Reviews’) and the lack of new studies in this area of research, a full systematic review of this topic was deemed not to be necessary for this thesis.

Two additional, relevant systematic reviews were identified. One reviewed the accuracy of guided intra-articular anaesthetic injections into the hip joint for diagnosing osteoarthritis (Dorleijn, Luijsterburg, Bierma-Zeinstra, & Bos, 2014) and the other (Smith, Hilton, Toms, Donell, & Hing, 2011) investigated the accuracy of MRA and magnetic resonance imaging (MRI) for diagnosing acetabular labral tears. Of the remaining publications identified by this search, 20 were diagnostic accuracy studies of various medical imaging or diagnostic procedures (12 for MRI/MRA, 2 for x-ray, and 5 for guided anaesthetic injections). A mixture of narrative reviews, expert commentary and studies that investigated associations (but not accuracy) between clinical findings and pathology made up the balance of the identified publications.

### **Identified systematic reviews**

An overview of the identified systematic reviews is presented in Table 5.2 below. The quality of these reviews was considered using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Liberati et al., 2009; Moher et al., 2009). The PRISMA guidelines were developed to help ensure clarity and transparency in the reporting of systematic reviews and is not intended as a tool to assess the quality of such reviews (Liberati et al., 2009). However, by considering each item in the guidelines and the relevant content provided by the authors of a systematic review, an appreciation of the quality of that review can be ascertained. Key conclusions resulting from this critique are presented in the following text.

**Table 5.2 Overview of systematic reviews of diagnostic accuracy studies investigating physical tests for intra-articular pathology of the hip**

Systematic Review	'Relevant' Pathology <sup>1</sup>	Number of included Studies	Meta-analysis performed
Leibold et al. 2008	Labral Lesions	16	No
Burgess et al. 2011	Labral Pathology	21	No
Tijssen et al. 2012	Labral Pathology FAI OA	1	No
Reiman et al. 2013	OA AVN Intra-articular pathology	25	Yes <sup>2</sup>
Rahman et al. 2013	Symptomatic OA	16	No
Reiman et al. 2014	FAI Labral tear	21	Yes <sup>2</sup>

<sup>1</sup> Some reviews included other pathologies not relevant to this thesis e.g. gluteal tendinopathy

<sup>2</sup> Performed for FADDIR and FIR tests

OA = osteoarthritis; FAI = femoroacetabular impingement; AVN = avascular necrosis

Two systematic reviews (Reiman et al., 2014a; Reiman et al., 2013) specifically followed the PRISMA guidelines. It is clear that both of these reviews are of high quality, as are two others (Burgess et al., 2011; Rahman et al., 2013). The two remaining reviews (Leibold et al., 2008; Tijssen et al., 2012) did not provide enough detail to enable the writer to determine the quality. All reviews provided an appropriate rationale, clear objectives and detail of the characteristics and results for individual included studies. Four of the six reviews (Burgess et al., 2011; Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012) used the Quality of Diagnostic Accuracy Studies (QUADAS) tool to assess the risk of bias in the individual studies included within their review (Whiting et al., 2006). Leibold et al. used the Standards for Reporting Studies of Diagnostic Accuracy (STARD) checklist, initially designed to improve the quality of reporting of diagnostic accuracy studies, but also used as a retrospective checklist to evaluate methodological quality (Bossuyt et al., 2003a). Rahman et al. examined the quality of all studies using predetermined internal and external validity criteria prior to inclusion in their review. Whilst these criteria are similar to those included in the QUADAS, they are not as extensive. Although these authors provided a list of excluded studies, they did not provide detail as to why these studies were excluded.

Other strengths common to all reviews except Leibold et al. were that they provided detail of the source of information, clear eligibility criteria and screening procedures for determining eligibility. Search strategies were well described for Reiman et al. (2014), Reiman et al. (2013), Rahman et al. (2013) and Tijssen et al. (2012). The most common failing in reporting across the reviews was insufficient detail regarding the process of data extraction and detail regarding the actual data collected, although this later information could be assumed by considering the characteristics and results tables that were provided. There was sufficient clarity and transparency in the reporting of four systematic reviews (Burgess et al., 2011; Rahman et al., 2013; Reiman et al., 2014a; Reiman et al., 2013) to allow a conclusion that these reviews were well conducted and represent ‘best-evidence’ at this point in time.

### **Studies identified that were not included in systematic reviews**

Three studies (Chong et al., 2013; O'Donnell et al., 2014a; Ochiai et al., 2011) that were not included in the abovementioned systematic reviews were identified by the current literature search. These studies, along with those in the Rahman review that were not examined for bias using a validated tool, were critiqued by the author using the QUADAS 2 checklist (Whiting et al., 2011). This tool defines the ‘quality’ of a study by determining both the risk of bias across four domains (patient selection, index test, reference standard, flow & timing) and any ‘concern’ regarding the applicability of study to the research question. Table 5.3. provides detail for the assessment of these studies. One study (O'Donnell et al., 2014a) was judged as ‘Low’ for both bias and applicability across all domains and therefore can be considered to have an overall “low” risk of bias and concern regarding applicability (Whiting et al., 2011). All of the other studies are at risk of bias and have concern regarding their applicability to the research question (detail provided in the following text). Evidence from these additional studies should be considered in the context of this assessment.

**Table 5.3 Summary of quality assessment of individual studies (QUADAS-2)**

Study	Risk of Bias				Applicability Concerns		
	Patient selection	Index Test	Reference test	Flow & Timing	Patient selection	Index Test	Reference test
Birrell	Low	High	Low	Unclear	Low	Unclear	Low
Chong	High	High	High	High	High	Unclear	High
Holla	High	Unclear	Unclear	High	High	Low	Low
Ochiai	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Low
O'Donnell	Low	Low	Low	Low	Low	Low	Low

Low, low risk of bias or concern regarding applicability; High, high risk of bias or concern regarding applicability; Unclear, risk of bias or concern regarding applicability is unclear.

### 5.2.1 Diagnostic accuracy of physical tests

The following sections provide a summary of the diagnostic accuracy of physical tests with respect to key intra-articular pathologies based on the evidence identified in the reviews and experimental papers.

#### **Acetabular labral pathology & FAI**

Three of the identified systematic reviews (Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012) considered and reported on diagnostic accuracy studies of physical tests to identify labral pathology and/or FAI. The two earlier reviews (Burgess et al., 2011; Leibold et al., 2008) only included studies that focussed on labral pathology. The most recent (and highest quality) review (Reiman et al., 2014a) identified 21 relevant studies which investigated a total of twelve individual tests. Of these, meta-analysis was appropriate for only the 'Flexion Adduction Internal Rotation' (FADDIR) and 'Flexion Internal Rotation' (FIR) tests. Pooled results for the FADDIR from four studies that used MRA as the reference standard were reported as: sensitivity 0.94 (95% CI 0.90 to 0.97), specificity 0.09 (95% CI 0.02 to 0.23), positive likelihood ratio 1.02 (95% CI 0.96 to 1.08), negative likelihood ratio 0.45 (95% CI 0.19 to 1.09) and diagnostic odds ratio (DOR) 5.71 (95% CI 0.84 to 38.86). Reiman and colleagues calculated the pre-test to post-test probability changes associated with this test. They reported that a positive FADDIR actually led to a small decline (1%) in the probability of the presence of FAI/labral pathology. A negative FADDIR also decreased the probability of this pathology being present (from 90% to 70%) although the confidence intervals around

the point estimate of the negative likelihood ratio for this test were very wide, creating doubt over the usefulness of this test.

With respect to the FIR test, pooled values demonstrate a high sensitivity [0.96 (95% CI 0.81 to 0.99)] suggesting that a negative test may have utility for screening for this pathology. However, the confidence intervals around the point estimate of the negative likelihood ratio were very wide and included the value 1, indicating that this test has limited clinical usefulness for ruling out such pathology. Pooled values for specificity [0.25 (95% CI 0.01 to 0.81)] and for the positive likelihood ratio [1.28 (95% CI 0.72 to 2.27)] indicate that a positive FIR test result is not useful clinically. Interestingly, despite drawing attention to the wide CI's associated with these tests, Reiman et al. concluded that the FADDIR and FIR tests have sufficient evidence to suggest that they may be useful as screening tests for FAI/labral pathology. However, they preface this by commenting that this conclusion is based on low quality studies that included patients with a high likelihood of having this pathology (i.e. confirmation bias).

None of the other systematic reviews performed meta-analyses of results for tests for labral tears or FAI. All reviews came to a similar conclusion with respect to 'ruling in' a diagnosis of labral pathology or FAI i.e. there was not any test specific enough to recommend its use for this purpose. However, based on the findings of one study (McCarthy & Busconi, 1995), Reiman et al. (2013) reported 'intriguing' findings for the Thomas test. McCarthy & Busconi reported sensitivity of 0.89 and specificity of 0.92 suggesting that this test could be used to both 'rule out' and 'rule' in labral pathology. However, these authors published two versions of this study, one with 59 participants and the other with 94. The data provided within these papers is unclear, making a correct determination of the accuracy of this test impossible. Reiman et al. did not elaborate on their reasons for describing this result as 'intriguing' although it seems that this might be the reason. Interestingly, Reiman and colleagues rated this as a 'High' quality study based on their critique (using the QUADAS-2 tool). Nevertheless, they were not confident enough to recommend that this evidence was convincing.

Leibold et al. (2008) and Burgess et al. (2011) concluded that there was sufficient evidence to support the negative results from 'physical tests' to rule out labral pathology. It seems that this conclusion was based purely on the high point estimates of sensitivity (most between 0.75 and 1.0) reported by the included studies. In contrast, Tijssen et al. (2012) stated that they were unable to make conclusive recommendations

regarding the clinical utility of physical tests, for either confirming or discarding a diagnosis of labral pathology, due to the low quality of the studies in their review. The other conclusion common to all of these reviews was that the quality of diagnostic accuracy studies to date is generally low, and that further good quality research was necessary.

The current literature search identified one relevant study (Ochiai et al., 2011) that was not included in any of the above systematic reviews. Unfortunately, only an abstract of this study could be retrieved, as it appears that a full text article has not been published. These authors described a new test (The Twist Test) for labral pathology that they had investigated in 247 patients using MRA as the reference standard. They reported sensitivity of 68% and specificity of 71%. Quality assessment of this study was not possible given the insufficient detail provided in the abstract.

### **Symptomatic osteoarthritis**

Two of the systematic reviews (Rahman et al., 2013; Reiman et al., 2013) considered and reported on diagnostic accuracy studies of physical tests to identify osteoarthritis. Rahman et al. only included one relevant study (Sutlive et al., 2008). This study reported that two variables that were associated with x-ray findings of OA i.e. ‘posterior pain whilst squatting’ [specificity 0.96 (95% CI 0.91, 0.99); positive LR of 6.1 (95% CI 1.5, 25.6)] and ‘groin pain with active abduction or adduction’ [specificity 0.94 (95% CI 0.89, 0.98); positive LR 5.7 (95% CI 1.7,18.6)]. Rahman and colleagues considered that these findings were ‘moderately useful’ for ‘ruling in’ a diagnosis of symptomatic OA. However, sensitivity was too low [0.24 (95% CI 0.09, 0.48) and 0.33 (95% CI 0.15, 0.57) respectively] and negative LR’s too large [(0.79 (95% CI 0.62, 1.0) and 0.71 (95% CI 0.52, 0.96)] to warrant using these variables to ‘rule out’ osteoarthritis.

Reiman et al. (2013) included two relevant studies (Sutlive et al., 2008; Youdas, Madson, & Hollman, 2010). Based on the study by Youdas et al., the authors of this review concluded that the only physical test that is useful for identifying people with hip OA was strength testing of the hip abductors. Youdas et al. reported that reduced hip abduction strength, when normalised to the individual’s body weight, was associated with hip osteoarthritis. They demonstrated that when hip abductor strength is <30% of body weight, the probability of the presence of OA shifts from 5% (pre-test) to 16% (post-test). However, the inclusion criteria and reference standard used in this study are of concern. Participants were volunteers with hip OA and associated “unilateral

impairment”, who were not currently “seeking treatment for hip impairment” (no mention of associated pain). Determination of OA status was based on self-reports i.e. people who reported “being told by their physician that they had a medical diagnosis OA”. Although these authors checked prospective participants medical records for x-ray evidence of degenerative hip disease, no detail was provided about the method of radiological classification used (if any). Consequently, there is very likely to be verification bias in this study and its findings may not be relevant to patients with hip pain.

Three other studies (Birrell et al., 2001; Chong et al., 2013; Holla et al., 2012) that examined the accuracy of physical tests for identifying OA were identified by the current literature search. Each of these studies was judged to be ‘at risk of bias’ and to have ‘concern regarding applicability’ (see Table 5.3). Birrell et al. and Holla et al. used radiographic evidence of OA as the reference standard and considered how accurately changes in range of motion predicted the presence of OA. Both studies concluded that restricted range of internal rotation was the best predictor of OA. Holla et al. reported that the probability of osteophytes or joint space narrowing being present on x-ray shifted from 25% pre-test to 46% post-test if internal rotation was 24 degrees or less [sensitivity 0.56 (95% CI 0.45, 0.67); specificity 0.78 (95% CI 0.72, 0.83); positive LR 2.55 (95% CI 1.86, 3.51)]. Birrell et al. reported that all patients in their study with OA had internal rotation of  $\leq 28$  degrees. However, they did not report sensitivity, specificity or LR values for this finding.

Chong et al. (2013) studied the relationship between radiographic evidence of OA in ten people with hip pain and their response to a fluoroscopy-guided injection of anaesthetic and steroid. They reported that eight patients had a 50% or greater reduction in pain as a result of the injection and that there was a statistically significant association between this magnitude of pain relief and the presence of pain on testing hip internal rotation prior to the injection. They also reported that there was no association between the presence or severity of radiographic OA and pain relief from the injection. Sensitivity, specificity or LR values were not reported.

Whilst there is a relatively consistent trend towards painful and restricted hip internal rotation being associated with radiographic OA across these studies, this association needs to be considered with caution given the overall quality of the evidence.

### **Ligamentum teres pathology**

None of the identified systematic reviews included studies that have investigated the diagnostic accuracy of tests for pathology affecting the ligamentum teres. However, one such study (O'Donnell et al., 2014a) was identified by the current review. Assessment of this study using the QUADAS-2 tool suggests that it has an overall 'low' risk of bias and 'low' concern regarding applicability. O'Donnell reported sensitivity of 0.90 (95% CI 0.39, 0.56) and a specificity of 0.85 (95% CI 0.75, 0.92) for a novel test ('The Ligamentum Teres Test') developed by the surgeons who published this paper. This test involves internal rotation followed by external rotation of the hip. It is performed with the hip positioned in approximately 70° of flexion and 30° "away from a fully abducted position". These authors define a positive test as pain provocation that occurs early in the range, of either internal or external rotation, provided that it is relieved by rotation in the opposite direction. It is proposed to place maximum tension on the ligament whilst avoiding any bony or soft tissue impingement. O'Donnell reported data in 2 x 2 tables enabling the calculation of likelihood ratios and diagnostic odds ratios. Based on this one study, this test appears to have diagnostic utility to both identify ligamentum teres pathology, with a positive LR of 6.5 (95% CI 3.7, 11.3), and to rule it out, with a negative LR of 0.11 (95% CI 0.05, 0.23).

### **Intra-articular pathology**

Two studies (Martin et al., 2008; Maslowski et al., 2010) have considered the diagnostic accuracy of physical tests to identify 'intra-articular' pathology rather than a specific structure within the joint. Both authors used fluoroscopy guided intra-articular injections of anaesthetic (along with a steroid) as their reference standard. Martin et al. described a positive anaesthetic response as a 50% or greater reduction in pain relief felt over the 2 hours that followed the procedure "when performing activities and getting into positions that in the past consistently aggravated their pain". Maslowski et al. required an 80% reduction in the pain score immediately post injection compared the score obtained at 'baseline'. Both of these studies were included in three of the abovementioned systematic reviews (Reiman et al., 2014a; Reiman et al., 2013; Tijssen et al., 2012) and both studies were assessed as 'moderate' quality by the authors of these reviews.

Martin et al. and Maslowski et al. investigated the accuracy of the 'Flexion Abduction External Rotation' (FABER) test. Martin et al. reported a sensitivity of 0.6 (95% CI 0.41, 0.77) for this test, whereas Maslowski et al. reported the higher value 0.82 (95%

CI 0.57, 0.96). Specificity was similar across these two studies (0.18 (95% CI 0.07, 0.39) and 0.25 (95% CI 0.09, 0.48) respectively). Martin et al. reported very poor positive (0.73) and negative (2.2) likelihood ratios. Although Maslowski and colleagues observed slightly better values (1.1 and 0.72), neither likelihood ratio suggests that these tests have significant clinical utility. Martin et al. also investigated the FADDIR test reporting sensitivity of 0.78 (95% CI 0.59, 0.89) and specificity of 0.10 (95% CI 0.03, 0.29). Maslowski et al. investigated the Scour (aka Quadrant) and FIR tests. Sensitivity for these tests was 0.5 (95% CI 0.26, 0.74) for Quadrant and 0.91 (95% CI 0.68, 1.0) for FIR. Specificity values were much lower at 0.29 (95% CI 0.12, 0.51) and 0.18 (95% CI 0.05, 0.40) respectively.

These studies provide some evidence that these tests are more sensitive than specific and that negative findings may strengthen a conclusion that intra-articular pathology is absent. However, this conclusion needs to be considered with caution given the low number of relevant studies, the 'moderate' quality of the studies and the width of the confidence intervals around the point estimates of accuracy.

### **5.3 Review Summary**

The authors of the abovementioned systematic reviews have recommended that there is a need for further, high quality studies to determine the diagnostic accuracy of physical tests for intra-articular hip joint pathology. The limitations and weaknesses of existing studies, be they in the conduct of the study or in the reporting of the study, make it difficult to draw sound conclusions about the diagnostic accuracy of hip joint tests. Based on the meta-analysis of Reiman et al. (2014a), the FADDIR and FIR tests may have some clinical utility as screening tests to rule out FAI/labral pathology. Two moderate quality studies (Martin et al., 2008; Maslowski et al., 2010) provide some evidence that the FADDIR, FIR and quadrant tests are sensitive for intra-articular pathology of the hip.

Reiman et al. (2013) concluded that a reduction in hip abductor strength was useful in identifying people with hip osteoarthritis based on the findings of one study (Youdas et al., 2010). A restriction in range of movement of internal range has been associated with osteoarthritis in a number of studies (Altman et al., 1991; Birrell et al., 2001; Holla et al., 2012). However, there are some concerns regarding bias in these studies making it difficult to make a definitive call on this finding. The ligamentum teres test appears to have diagnostic utility for pathology of this ligament with one study (O'Donnell et al.,

2014b) demonstrating that it has both high sensitivity and specificity. On the basis of the evidence considered in this review, it is clear that additional, high quality research is required to provide conclusive evidence regarding the diagnostic utility of physical tests in assessment of the painful hip.

## **5.4 Methodological considerations**

There are a number of methodological factors that affect the internal and external validity of diagnostic accuracy studies. The following section summarises key considerations crucial to the design, conduct and analysis of such studies, with reference to the studies identified in the current review.

### **5.4.1 Study design & patient spectrum**

Essentially, DA studies determine the ability of one or more tests (the ‘index’ tests) to predict the presence or absence of a specific ‘disease’ or pathology. To achieve this, the results of the index test are compared to those of a ‘gold’ or ‘best available’ standard (aka the ‘reference’ test) that establishes the true presence or absence of this pathology. There is a consensus of opinion that the ideal design should be a prospective cohort design where consecutive patients from a relevant clinical population are evaluated with the index test and the reference test (Fritz & Wainner, 2001; Jaeschke, Guyatt, & Sackett, 1994b; Lijmer et al., 1999). Only four of the studies identified by the current literature search were clearly prospective cohort studies with consecutive patients (Ayeni et al., 2014a; Birrell et al., 2001; Narvani et al., 2003; O'Donnell et al., 2014a). Whilst other studies may have been of such design, there was insufficient detail reported to accurately determine if this was the case.

A case-control design was used by two studies (Verrall, Slavotinek, Barnes, & Fon, 2005; Youdas et al., 2010). This design typically creates spectrum bias due to the fact that cases are retrospectively selected (creating the possibility that only the more obvious or easily diagnosed cases are included and more subtle cases excluded) and that controls (people highly likely not to have the condition of interest) are not representative of the population that would normally be tested with the index test. This distorts both clinical presentation and the prevalence of the condition in the cohort being studied and can cause over-estimation of both the sensitivity and specificity of included tests (Fritz & Wainner, 2001). This point was highlighted by Lijmer et al. (1999) who evaluated the results of 193 published diagnostic accuracy studies and compared estimates of diagnostic accuracy (of a wide range of medical tests) reported in high

quality studies with those from lower quality studies. These authors reported that of all the shortcomings in design and conduct of the diagnostic accuracy studies that they evaluated, the case-control design had the single largest effect on results (over-estimation of the diagnostic odds ratio by a factor of three). Whilst a prospective design is recommended (Bossuyt et al., 2003b; Fritz & Wainner, 2001; Jaeschke et al., 1994b), Lijmer and colleagues reported that retrospective data collection did *not* generate different estimates of diagnostic accuracy than the studies that collected data prospectively. Seven studies included in the current literature review were retrospective (Burnett et al., 2006; Byrd & Jones, 2004a; Chong et al., 2013; Holla et al., 2012; Myrick & Nissen, 2013; Ochiai et al., 2011; Wang, Yue, Zhang, Hong, & Li, 2011).

Recruitment of consecutive patients presenting with symptoms and signs that are consistent with the condition of interest is considered important to limit the bias created by populations where cases are selected. However, consecutive recruitment is unfeasible for some studies. For example, with regard to the hip joint, the acknowledged ‘gold’ standard is arthroscopy, but, it would be unnecessary and inappropriate for every patient presenting at a primary medical care facility (e.g. general practice or physiotherapy clinic) to undergo this procedure and to be exposed to the attendant risks. An alternative method of recruitment is to include consecutive patients who meet the criteria for a specific diagnostic investigation that would be an appropriate reference standard for the condition of interest (e.g. MRI, MRA, guided intra-articular injections of anaesthetic or arthroscopy). A benefit of selecting such a cohort is that it helps to ensure that all participants have the same reference test, therefore avoiding verification bias (Bossuyt et al., 2003a; Simel, Rennie, & Bossuyt, 2008). The drawback with this solution is that the findings of studies that have selected participants on this basis are only relevant to patients that meet the same criteria as those selected for the study. In particular, when arthroscopy is the chosen reference standard, the spectrum of patients becomes quite narrow i.e. only those with symptoms/signs severe enough to warrant surgical intervention. A number of the studies that investigated the accuracy of clinical tests of the hip identified in the current review included patients who had undergone arthroscopy (Myrick & Nissen, 2013; O'Donnell et al., 2014a; Philippon, Maxwell, Johnston, Schenker, & Briggs, 2007; Suenaga et al., 2002; Wang et al., 2011).

#### **5.4.2 Reference test**

Ideally, this test would be the ‘gold standard’ test and would have 100% sensitivity and specificity for the condition of interest. However, there are few tests in medicine that

meet this stringent standard. In the absence of a perfectly accurate gold standard, a test “that is considered to be the best available under reasonable conditions” should be utilised (Versi, 1992). To avoid verification bias and consequent over-estimation of diagnostic accuracy, all participants should undergo the same reference test or combination of reference tests (Fritz & Wainner, 2001; Lijmer et al., 1999).

A variety of reference tests were used in the studies identified in the current review. Three used fluoroscopy guided anaesthetic injections (Chong et al., 2013; Martin et al., 2008; Maslowski et al., 2010), three used MRI (Hananouchi et al., 2012; Joe et al., 2002; Verrall et al., 2005), four used X-ray (Birrell et al., 2001; Holla et al., 2012; Sutlive et al., 2008; Youdas et al., 2010), four used MRA (Ayeni et al., 2014a; Narvani et al., 2003; Ochiai et al., 2011; Troelsen et al., 2009), and ten used arthroscopy as the reference standard (Burnett et al., 2006; Byrd & Jones, 2004a; Clohisy et al., 2009; Mitchell et al., 2003; Myrick & Nissen, 2013; O'Donnell et al., 2014a; Philippon et al., 2007; Springer et al., 2009; Suenaga et al., 2002; Wang et al., 2011).

### **Arthroscopy**

Whilst many authors state that arthroscopy is considered the ‘gold’ or best reference test for intra-articular hip joint disorders, patients who have undergone an arthroscopy are typically those that have had their hip pain for a long time and who most likely have pathology that was considered by the surgeon to be treatable by arthroscopy (Groh & Herrera, 2009; Keeney et al., 2004; Mitchell et al., 2003). Arthroscopy is expensive, technically demanding and not without risk (Harris et al., 2013). It is ideally reserved for patients with clinical and imaging findings that strongly suggest intra-articular pathology (excluding significant arthritis), with persistent pain and who have failed conservative treatment. Consequently, the findings of studies that have used surgery as the reference standard are biased towards that spectrum of patients and are unlikely to be as applicable to patients who have not seen an orthopaedic surgeon and who have a much lower suspicion of intra-articular pathology (i.e. a lower pre-test probability). As such, the findings of such studies should only be generalised to a wider group of patients with caution.

It seems that the rationale for using surgical findings as the reference standard is that if a surgeon can see pathological tissue, then that tissue must be the cause of the patient’s pain. Whilst this makes some sense, it contradicts the large body of evidence that has demonstrated that pathological changes are present in large percentages of people who

are (and always have been) asymptomatic. This has been shown to be the case in multiple regions of the body including the hip (Abe et al., 2000; Jung et al., 2011; Register et al., 2012; Silvis et al., 2011), the lumbar spine (Boos et al., 2000; Brinjikji et al., 2015) and the shoulder (Bonsell et al., 2000; Connor, Banks, Tyson, Coumas, & D'Alessandro, 2003).

Another factor to consider is the reliability of surgical identification and grading of abnormal pathology (Fritz & Wainner, 2001). The inter-observer agreement of surgeons' ability to categorise tissue damage identified during surgery has been examined in a number of studies (Kuhn et al., 2007; Nepple et al., 2012; Spahn, Klinger, Baums, Pinkepank, & Hofmann, 2011). Kappa values ranging between 0.19 and 0.80 have been reported. Specific to the hip, Nepple et al. reported that the reliability of experienced surgeons in determining whether or not the labrum was normal was between 0.36 and 0.84 (average 0.69). Whilst these levels of reliability might be considered acceptable for clinical use, they are of concern if surgical findings are to be used as the reference standard for diagnostic accuracy studies. The possibility of misclassification bias cannot be discounted when surgical findings are used as the reference standard.

### **MRI or MR arthrography**

MRI and MRA were used in seven of the studies identified in the current review (see detail above). There is some evidence that a field strength of 3 Tesla (T) is superior to 1.5T MRI for imaging of musculoskeletal pathology of the fingers, wrist and hand. This issue does not appear to have been investigated for the hip joint (Schoth et al., 2008; Wieners et al., 2007). Of the studies in the current review, one used a 3T field strength (Hananouchi et al., 2012), two used 1.5T (Ayesi et al., 2014a; Troelsen et al., 2009), one used 1.0T (Narvani et al., 2003). Three studies did not report these details (Joe et al., 2002; Ochiai et al., 2011; Verrall et al., 2005).

The reliability and diagnostic accuracy of findings identified by MRI and MRA needs to be considered. Silvis et al. (2011) investigated the inter-rater and intra-rater reliability of *experienced* musculoskeletal radiologists reporting of pathological changes identified via MRI (3 Tesla). The authors reported that kappa values for inter-rater reliability were only 0.37 for hip osteochondral lesions and 0.41 for labral tears. Similarly, intra-rater reliability for these pathologies was low (kappa values of 0.37 and 0.42 respectively). Contrasting results were reported by Kumar et al. (2013) who investigated the reliability

of experienced musculoskeletal radiologists using MRI (3 Tesla) for the grading of four radiological features of OA i.e. cartilage defects, bone marrow edema-type lesions (BMELs), subchondral cysts and labral tears. Kumar and colleagues reported substantial kappa values for intra-rater reliability (of 0.70, 0.79, 0.78 and 0.73 respectively) for these pathologies. Inter-rater reliability values were much lower for cartilage defects (0.57) and BMELs (0.55) although acceptable for cysts (0.71) and labral tears (0.65).

A systematic review and meta-analysis (Smith et al., 2011) of the diagnostic accuracy of MR arthrogram imaging (compared to findings from arthroscopy as the reference standard) reported a pooled sensitivity of 0.87 (95% CI: 0.84 to 0.90) and pooled specificity of 0.64 (95% CI 0.54, 0.74) (see page 213 for full details of this review). Smith et al. concluded that MRA is useful for identifying labral tears of the hip, but should not be relied upon as a stand-alone diagnostic test. The sensitivity and specificity for MRI and MRA for the identification of tears of the ligamentum teres has also been investigated. One study (Datir et al., 2014) has demonstrated that the accuracy of 3T MRA and MRI are essentially the same for complete tears of this ligament (sensitivity of both MRI and MRA was 67%, specificity of MRA was 100% and that for MRI was 99%). In respect to partial tears, Datir and colleagues reported that MRA has a higher sensitivity (0.83 versus 0.41) and specificity (0.93 v 0.75) than MRI (see Chapter 7, page 216 for further detail regarding this study). In contrast to these findings, Devitt et al. (2014) reported higher sensitivity (0.91) and much lower specificity (0.09) for partial tears of this ligament (using 3T MRI). It seems likely that these contrasting results reflect differences in the classification criteria and imaging protocols used in these two studies. Devitt and colleagues also reported sensitivity of 0.78 and specificity of 0.32 for the identification of hypertrophic ligamentum teres (see Chapter 7, page 216 for further detail).

McGuire et al. (2012) considered the diagnostic accuracy of 1.5T MRI compared to arthroscopy and the differences in accuracy between radiologists who have sub-specialized in musculoskeletal radiology and less experienced 'general' radiologists (see page 214 for further detail). Not surprisingly, these authors reported that those who have specialized in musculoskeletal radiology are substantially more accurate in the reporting of labral tears, chondral damage and FAI (overall accuracy 85%, 69% and 82% respectively) than the general radiologist (70%, 40% and 59%).

Finally, a high prevalence of pathology identified by MRI and MRA has been demonstrated in asymptomatic populations, suggesting that the presence of pathology identified by MRI/MRA in patients with hip pain, does not confirm that the source of a patient's pain has been established (Frank et al., 2015; Kwee, Kavanagh, & Adriaensen, 2013) (see page 206 for further details).

### **Radiology**

Given that there is no diagnostic test for OA, a diagnosis of OA has classically been made on the basis of the ACR criteria that combine clinical and radiological features (Altman et al., 1991; Bijlsma, Berenbaum, & Lafeber, 2011). However, the reliability of the ACR criteria has been demonstrated to be poor (Reijman, Hazes, Koes, Verhagen, & Bierma-Zeinstra, 2004a). A number of alternative methods of using radiological findings to define and grade OA have been described (Croft, Cooper, Wickham, & Coggon, 1990; Kellgren & Lawrence, 1957; Lane, Nevitt, Genant, & Hochberg, 1993). A systematic review that explored the validity, reliability and applicability of a wide range of definitions concluded that the intra and inter-tester reliability of the minimal joint space, Kellgren & Lawrence and the Lane Index were "good" (kappa values in most of the included studies between 0.7 to 0.85), whilst the inter-rater reliability of the Croft and ACR criteria were relatively low (Reijman et al., 2004a). These reviewers reported that the minimal joint space criteria demonstrated the highest relationship with hip pain. However, they commented that the validity of these criteria has hardly been investigated. A more recent systematic review has demonstrated that the prevalence of osteoarthritis is higher when based on radiological definitions than that based on other definitions, including 'symptomatic OA' (where the patient has both radiological evidence and pain) and 'self-reported OA' (where patients report that they have previously been diagnosed with OA) (Pereira et al., 2011).

Four studies in the current review examined the diagnostic accuracy of physical tests of the hip for the identification of hip osteoarthritis using radiological imaging as the reference standard (Birrell et al., 2001; Holla et al., 2012; Sutlive et al., 2008; Youdas et al., 2010). Of these, all but one (Youdas et al., 2008) recruited people with both radiological evidence of OA and associated hip pain. Whilst each of these studies used x-ray as the reference standard, a variety of radiological grading methods were employed. Birrell et al. employed the method of Croft, a system reported by Reijman et al. (2004a) to have low reliability. Sutlive et al. using the more reliable Kellgren and Lawrence method and Holla et al. utilised the Altman & Gold grading system (Altman

& Gold, 2007), the reliability of which does not appear to have been determined. As discussed previously, Youdas et al. appeared not to use any grading criteria, relying solely on radiological evidence of OA recorded in their participant's medical records. Each of these different criteria focus on different radiological features and there seems not to be any consensus in regard to the choice of grading methods for hip OA across the diagnostic accuracy literature. This inconsistency makes it difficult to compare the findings from the various studies, and to make valid conclusions based on the results of the studies that did not use methods proven to be reliable.

Even if a valid and reliable grading method is employed, there are other factors to consider in regard to the appropriateness of using radiological evidence of hip OA as a reference standard. Firstly, a large percentage (56 to 76%) of patients with arthroscopically confirmed OA have normal radiographs suggesting that x-rays are not as sensitive as arthroscopy (McCarthy & Busconi, 1995). Similarly, there seems to be a poor relationship between radiological findings and pain or loss of function, suggesting that radiological evidence of OA does not prove that OA is causing any hip pain/impairment or that the degree of degenerative change determines the severity of symptoms (Birrell, Lunt, Macfarlane, & Silman, 2005; Kumar et al., 2013; Reijman et al., 2004b). Finally, x-ray does not provide useful information in regard to any soft-tissue causes of hip pain; thus, many common pathologies (such as labral tears, ligamentum teres tears, synovitis) will not be identified.

### **Intra-articular injection of anaesthetic**

Pain originating from intra-articular structures of the hip is commonly felt in the groin, buttock and thigh (Arnold et al., 2011; Leshner et al., 2008). However, it is well documented that pain from the lumbar spine and sacroiliac joint may be referred to these same areas (Bogduk, 2009; Young & Aprill, 2000). Determining if the source of such pain is from within the hip or not can be difficult, particularly when the patient has clinical signs (e.g. pain reproduced by hip tests and by lumbar spine tests) or imaging findings (e.g. degenerative changes in both the hip and the lumbar spine) that suggest that either source could be the source of pain. Fluoroscopy or ultrasound guided anaesthetic injections into the hip joint are commonly used in clinical practice to help make this differentiation (Arnold et al., 2011; Byrd & Jones, 2004a; Domb et al., 2009; Kivlan et al., 2011; Rho, Mautner, Nichols, & Kennedy, 2013). A significant reduction in pain (i.e. a positive anaesthetic response) reported by the patient after such a procedure suggests intra-articular pathology (Bogduk, 2004a, 2004b). Such injections

have been widely used as the reference standard for diagnostic accuracy studies of the sacroiliac and lumbar facet joints (Bogduk, 2004a; Dreyfuss, Michaelsen, Pauza, McLarty, & Bogduk, 1996; Laslett et al., 2005; Young & Aprill, 2000; Young, Aprill, & Laslett, 2003). Although the actual structure itself cannot be identified with this procedure, this is not of concern in the early stages of the clinical examination where initial hypotheses are being explored. The ability to ‘rule out’ an intra-articular source through a clinical examination has several benefits. In the case where the site of the pain allows for the possibility that the lumbar spine, sacroiliac joint, pubic symphysis or extra-articular hip joint structure could be the source of symptoms, ruling out intra-articular hip structures allows the examination to focus on these other regions. This would not only facilitate earlier identification of the actual origin of such pain, but would mean that unnecessary investigations of the hip joint would not be undertaken.

Dorleijn et al. (2014) performed a systematic review and meta-analysis of case series studies that had examined the diagnostic accuracy of guided anaesthetic injections (GAI) in patients who had pain that could not, on the basis of the clinical examination, be proven to originate from within the hip. Participants in these studies underwent an intra-articular GAI into the hip and subsequent total hip arthroplasty (THA) for end stage hip OA (Dorleijn et al., 2014). Pain relief after the THA was considered the reference standard and the accuracy of the GAI was determined by evaluating its ability to predict this outcome. The authors of this review reported a pooled sensitivity of 0.96 (95% CI 0.87, 0.99) and specificity of 0.42 (95% CI 0.09, 0.84) and concluded that clinicians could “cautiously” predict that patients with a negative anaesthetic response (NAR) would be less likely to get relief from a THA (negative LR of 0.09) than those that had a positive anaesthetic response (PAR). Whilst this suggests that a NAR is evidence that the hip joint is not the source of pain, the Dorleijn et al. review demonstrates that there is insufficient evidence to conclude that a PAR is diagnostic of hip OA.

Byrd and Jones (2004a) retrospectively compared the diagnostic accuracy of MRI, MRA and intra-articular injections to arthroscopy and reported an accuracy of 90% for intra-articular injections. Data provided by these authors enabled calculation of sensitivity [0.92 (95% CI 0.78, 0.97)] and specificity [0.33 (95% CI 0, 0.96)]. These findings are consistent with those of the abovementioned systematic review. A more recent study (Ashok, Sivan, Tafazal, & Sell, 2009) prospectively examined the diagnostic accuracy of FGAI in 48 consecutive patients with both hip *and* back pain. In

this study a PAR was considered to be  $> 70\%$  decrease in pain intensity. Of the 37 patients who had a PAR, 34 subsequently underwent THA. Thirty-three of these patients had complete relief of their pain following this procedure with the remaining patient reporting no relief. Eleven patients had a negative anaesthetic response (NAR). Ten of these underwent treatment of their lumbar spine and reported the ‘satisfactory’ relief. One NAR patient had a successful outcome following THA. Ashok et al. reported similar sensitivity (97%) to the previous authors but a much higher specificity (91%). This higher specificity is likely to reflect the fact that this study included some participants (most of those who had a NAR) that did not undergo THA. The reference standard for these patients was treatment of their lumbar spine. The different reference standards for PAR and NAR participants decreases the strength of the findings of this study.

Three studies in the current review used the response to fluoroscopy guided anaesthetic injections (FGAI) as the reference standard for diagnostic accuracy studies of physical tests for the hip (Chong et al., 2013; Martin et al., 2008; Maslowski et al., 2010). Differences in the cut-off point used to define a positive anaesthetic response (PAR) to these injections between these studies is readily apparent with two (Chong et al., 2013; Martin et al., 2008) requiring a 50% or greater reduction of pain intensity and the other (Maslowski et al., 2010) requiring an 80% reduction. This inconsistency is seen throughout the diagnostic accuracy literature, regardless of the joint of interest (including the shoulder, sacroiliac joint and lumbar spine). Descriptive definitions of a PAR such as ‘marked improvement’ or ‘near-complete’ relief have also been employed (Chronopoulos, Kim, Park, Ashenbrenner, & McFarland, 2004) but as these two examples demonstrate, wording has not been consistent.

One study (Kivlan et al., 2011), has examined the relationship between the percent relief reported by patients undergoing GAI and the pathological findings identified by arthroscopy. This study adds another dimension to the issue of setting cut-off points to define a PAR. These authors demonstrated that subjects with chondral pathology had greater relief from GAI (around 90% reduction) than those that did not have such pathology. They also observed that the presence of extra-articular pathology did not influence the percent relief in patients who had intra-articular pathology. Similarly, the presence of labral pathology or FAI had no consistent effect on the percent relief experienced by patients with these pathologies (reporting anywhere between 10% to 100% relief). These authors also reported that 80% of the subjects who had labral tears

or FAI identified by surgery, had less than a 50% anaesthetic response. Kivlan and colleagues concluded that abnormalities of the labrum without associated chondral pathology are unlikely to be symptomatic.

Adding to the mix that makes it difficult to compare findings of different studies is the method used to determine both the baseline pain intensity (pre GAI) and post procedure pain intensity. Of the three relevant studies in the current review, only Martin et al. (2008) provided this detail, reporting that participants were required to determine the degree of pain relief in the 2 hours post injection “while performing activities and getting into positions that in the past aggravate the pain”. Both Chong et al. (2013) and Maslowski et al. (2010) reported that pain intensity was scored before and after the injection however neither mentioned whether or not the patient was subjected to any clinical tests or asked to perform functional activities that were painful either prior to or after the injection.

To help address these issues in respect to the use of diagnostic injections in the spine and sacroiliac joint, the International Spine Intervention Society (ISIL) have developed guidelines for their members (Bogduk, 2004b). These guidelines suggest that the level of baseline pain needs to be of sufficient intensity that any change after GAI is ‘credible and meaningful’ and recommend a minimum of 20mm on a 100mm visual analogue scale (VAS). It is recommended that an observer who is independent to the person who administers the injection should evaluate the response, using validated instruments, both ‘immediately’ and for ‘some time’ after the block. This evaluation should include requiring the patient to perform activities that usually “are impeded or prevented by their pain”. The guidelines state that 100% relief of pain is the ideal, but that this might not be realistic if the patient has more than one source of pain and/or if there is some discomfort related to the procedure itself. They suggest that a response of < 50% should be considered a NAR, one between 50% and 74% as equivocal and 75% or greater as a PAR. These suggestions appear to be based on expert opinion rather than research evidence. Whilst these recommendations were made with respect to spinal diagnostic injections, there seems to not be any reason why they are not applicable to peripheral joint injections. However, it should be noted that the operating characteristics of a test will be influenced by the level of response selected (Griner et al., 1981). Lower cut-off point will increase sensitivity at the expense of specificity and vice versa.

Relevant to this discussion is a question over the reliability of patients in determining the intensity of the pain that they feel during provocative activities and/or the application of physical tests to their symptomatic hip. This important issue was explored in Chapter 3 (see page 76 for relevant discussion). Without evidence that patients can reliably report pain intensity over a period of time, the validity of utilising changes in pain intensity as a marker of success or failure following GAI is questionable. Our results demonstrated that both the within and between-session reliability of patient reports of pain intensity provoked by physical tests at the hip using the NPRS sufficient reliability to be clinically useful (White, McNair, Laslett, & Hing, 2015).

It has been reported that false positives from GAI may result when there is leakage of the anaesthetic from within the hip joint capsule to surrounding extra-articular structures (e.g. the iliopsoas bursa or tendon) (Mitchell et al., 2003). Whilst ultrasound or fluoroscopic guidance might ensure that the anaesthetic is placed within the joint, the addition of gadolinium to the injectate allows the radiologist to identify any leakage into the surround bursa or soft tissues (Mitchell et al., 2003). One drawback with the use of gadolinium contrast is that it can cause a temporary increase in joint pain (Mosimann et al., 2012; Saupe et al., 2009). Saupe and colleagues measured baseline pain scores in 294 patients undergoing MRA of the hip and compared these scores to reports of pain intensity 4 hours, 1 day and 1 week after the injection of gadolinium contrast. They reported mean baseline pain scores of  $2.5 \pm 2.5$  (measured via the visual analogue scale) and that there was no significant change in intensity immediately after the injection. However, four hours after this injection, an increase in intensity of  $1.2 \pm 1.7$  points was noted. Saupe et al. commented that this time frame was compatible with the development of a chemical synovitis that might be secondary to irritation of the synovium by the contrast. Alternatively, they suggested that it could also be related to the wearing off of the local anaesthetic they used (mepivacaine hydrochloride 2%). This increase in pain associated with the use of gadolinium needs to be considered when the pain response to GAI is used as a reference standard. Increased pain as a result of the gadolinium may lead to a false negative anaesthetic response and the conclusion that there is no intra-articular pathology. In the study by Kivlan et al. (2011), participants were reassessed within 2 hours of having the injection so that this flare at 4 hours was avoided.

It is worth noting that Saupe et al. utilised small volumes (mean  $0.9 \pm 0.3$  mL) of a relatively short-lasting anaesthetic to anaesthetise the “skin, joint capsule, and joint space” so that the injection of the contrast was not painful. When intra-articular anaesthetic is injected diagnostically (to help to differentiate intra-articular from extra-articular pathology), larger volumes are used and frequently, both a short-lasting (2-6 mL of Lidocaine 1%) and long-lasting (3.5 mL to 10 mL of Bupivacaine 0.5%) anaesthetic is included (Ashok et al., 2009; Byrd & Jones, 2004a; Deshmukh et al., 2010; Kivlan et al., 2011). Consequently, a larger dose of anaesthetic is administered and the post injection pain associated with gadolinium reported by Saupe et al. is unlikely to be an issue. Evidence to support this has been provided by Mosimann et al. (2012) who investigated the influence short and long-lasting anaesthetics on the post injection pain associated with hip arthrograms. These authors compared an injectate of gadolinium (0.1 mL), adrenalin (0.05 mL 1%) and 10mL Omnipaque 300 to one that also contained either 4 mL of bupivacaine (0.25%) or 4 mL of lidocaine (1%). They reported that the addition of bupivacaine eliminated any immediate post injection discomfort whereas those with no anaesthetic had a mean pain increase on the VAS of 0.96. Similarly, the pain intensity four hours post injection was only  $0.29 \pm 0.75$  (units on the VAS) compared to  $1.6 \pm 0.83$  for those without bupivacaine.

### **5.4.3 Index test description**

The findings of studies that have investigated the diagnostic accuracy of a test can only be generalised to a clinical setting if the authors of that study provide a detailed description of how the index test is performed and interpreted (Fritz & Wainner, 2001; Whiting et al., 2003). Several studies in the current review provided sufficient detail to allow replication of the tests they examined although the majority did not do so.

Examination of the descriptions provided for several commonly used clinical tests (including quadrant, FABER, ‘impingement’, log roll) reveals that the performance and ‘scoring’ of these tests is not standard across this body of research. For example, in respect to the FABER test, Martin et al. (2008) placed the participants foot “just proximal to the uninvolved knee” and a positive test was ‘reproduction of pain in any location’. Maslowski et al. (2010) placed the foot ‘on the knee’ of the uninvolved leg and defined a positive test as one that ‘recreate(s) the subjects pain’. Another study (Philippon et al., 2007) described a positive FABER test as “any loss of ROM compared to the unaffected hip”. Any comparison of the findings from such studies needs to take into consideration variation in test performance and/or interpretation.

#### **5.4.4 Time between index test and reference standard**

It is important that the time period between the application of the index test(s) and the reference test is as short as possible so that there is a reduced risk of any change in status of the condition of interest. It is possible that the underlying pathology may improve or deteriorate over time leading to ‘disease progression bias’ (Whiting et al., 2003). The results of any test will vary if that test is applied at different stages of the condition. This is less of a concern with chronic, stable conditions but is still important to consider. Most of the existing studies that have evaluated the accuracy of physical tests for identifying intra-articular pathology of the hip did not provide detail in this regard (Byrd & Jones, 2004a; Chong et al., 2013; Holla et al., 2012; Joe et al., 2002; Martin et al., 2008; Springer et al., 2009; Suenaga et al., 2002; Troelsen et al., 2009; Verrall et al., 2005; Wang et al., 2011). Three studies (Maslowski et al., 2010; O'Donnell et al., 2014a; Sutlive et al., 2008) reported performing the index test(s) immediately prior to the reference standard. Conversely, three studies (Myrick & Nissen, 2013; Narvani et al., 2003; Philippon, Briggs, Yen, & Kuppersmith, 2009) indicated that there were significant delays (4 weeks up to 2 years) after baseline testing before the reference test was employed.

#### **5.4.5 Blinding**

Interpretation of the index test should be made in the absence of knowledge of the results of the reference standard (and vice versa) to reduce the possibility of review (aka information) bias (Fritz & Wainner, 2001; Whiting et al., 2003). Similarly, knowledge of the participant's clinical presentation should be limited to the information that would normally be known in the clinical situation where the test is likely to be employed (Whiting et al., 2003). The performance and/or interpretation of the index test may be influenced by a researcher who is privy to additional information (e.g. the results of a previous MRI or X-ray) leading to over-estimation of measures of diagnostic accuracy.

Of the studies in the current literature search, only seven (Holla et al., 2012; Joe et al., 2002; Springer et al., 2009; Suenaga et al., 2002; Sutlive et al., 2008; Troelsen et al., 2009; Verrall et al., 2005) stated clearly that the index and reference tests were performed in a blinded manner. Two studies (Byrd & Jones, 2004a; Maslowski et al., 2010) reported that they did not utilise blind assessors. The remaining studies were either unclear or did not report this information.

#### 5.4.6 Sample size analysis

A small sample size in a diagnostic accuracy study is likely to result in imprecise estimates of sensitivity and specificity (Bachmann, Puhan, Riet, & Bossuyt, 2006; Fenn Buderer, 1996). Such imprecision will be reflected in wide confidence intervals around the point estimates, making interpretation of results and their application to clinical practice difficult. Calculation of sample size prior to the conduct of a diagnostic accuracy study enables the researcher to be more confident that an appropriate number of subjects will be included and hence, achieve more precise estimates of measures of diagnostic accuracy. Sample size calculations should take into account the minimum level of sensitivity and/or specificity that would be acceptable for the test(s) under investigation to be employed appropriately in clinical practice (Bachmann et al., 2006; Fenn Buderer, 1996; Jones, Carley, & Harrison, 2003). Another important consideration is the effect that the prevalence of the target condition has on the precision of these estimates of accuracy. Consequently, the presumed prevalence should be included in any sample size calculations (Bachmann et al., 2006; Fenn Buderer, 1996; Jones et al., 2003).

Only one study (Maslowski et al., 2010) identified by the current search has reported performing a sample size analysis. However, these authors did not provide any details regarding acceptable levels of sensitivity or specificity or of expected prevalence of the target condition. The lack of consideration of this important component of study design is not confined to the musculoskeletal literature. A recent review (Bachmann et al., 2006) of all the diagnostic accuracy articles published in high quality medical journals (including *BMJ*, *Lancet* and *JAMA*) reported that only 5% reported an a priori sample size calculation.

For the studies that investigated the diagnostic accuracy of physical tests for the hip (identified by the current search), sample sizes ranged between 18 (Narvani et al., 2003; Troelsen et al., 2009) and 344 (Holla et al., 2012). The majority of studies included between 40 and 76 participants. Sample size in retrospective studies has typically been determined by the number of patients that have had the reference standard (e.g. arthroscopy, MRI/MRA) in the course of their management and for whom adequate details regarding the index test have been recorded in the clinical notes. Of the four studies that had 100 or more participants (Holla et al., 2012; Joe et al., 2002; Myrick & Nissen, 2013; Ochiai et al., 2011), three were retrospective studies.

## **5.5 Methods and procedures of the current study**

The methodological issues discussed in the preceding review were considered in the design and conduct of the current study. Results of this study have been reported in a manner consistent with relevant guidelines and critiquing tools for diagnostic accuracy studies (Bossuyt et al., 2003a; Bossuyt et al., 2003b; Whiting et al., 2003; Whiting et al., 2011)

### **5.5.1 Study design**

This study was an analytic, non-experimental, prospective cohort study designed to determine the diagnostic accuracy of individual components of the clinical examination.

### **5.5.2 Sample size**

A power and sample size estimation was performed employing the method recommended by Jones et al. (2003) and using the following parameters: expected levels of sensitivity and specificity = 90%; acceptable width of confidence intervals (either side of the point estimate) = 15%; the probability that lower confidence limit will fall within these limits = 5% and the likely prevalence of a positive response to intra-articular anaesthetic = 50%. The value of these parameters were determined after consideration of the findings of similar studies (Martin et al., 2008; Maslowski et al., 2010) and on expert opinion (Childs & Cleland, 2006). This calculation indicated that a sample size of 62 was necessary to determine the diagnostic accuracy of hip joint tests with the degree of precision required. The final sample size was increased to 70 to allow for possibility that some participants might not complete the protocol. This sample size is in line with existing studies that have used similar methods and analysis as the proposed study (Flynn et al., 2002; Hicks, Fritz, Delitto, & McGill, 2005; Sutlive et al., 2008; Vicenzino, Collins, Cleland, & McPoil, 2010; Vicenzino, Smith, Cleland, & Bisset, 2009). Sample size in these studies range from 42-71 subjects.

### **5.5.3 Participants**

Between June 2012 and December 2013, consecutive patients presenting at the referring medical specialist's (see Referring Specialists below) clinic, who met the following inclusion and exclusion criteria, were invited to participate in the study. The primary inclusion criteria for this study was that the patient's symptoms and signs were consistent (in the opinion of the specialist) with those of intra-articular pathology of the hip and that a magnetic resonance imaging arthrogram (MRA) with fluoroscopy guided anaesthetic injection (FGAI) was needed as part of the patient's diagnostic work-up.

Also, participants had to be aged between 15 and 80 years old and have unilateral pain in the groin or deep buttock region that had been present for a minimum of one month at the time of data collection. Ethical approval for this study was granted by the New Zealand Ministry of Health, Northern X Regional Ethics Committee (Reference number NTX/11/07/066) (See Appendix 9).

Exclusion criteria were patients who:

- had undergone previous hip joint surgery or;
- had undergone treatment for low back pain within the previous 12 months or;
- had evidence of lumbar radiculopathy, systemic disease or illness or;
- were claustrophobic or;
- were pregnant or;
- had severe pain (greater than 9 on the NPRS) or;
- were unable to speak English or;
- had any contraindication undergoing the MRA/FGAI procedure (as determined by the radiologist performing the procedure).

#### **5.5.4 Referring specialists**

Two sports physicians and one orthopaedic surgeon (all recognised nationally for their expertise in the diagnosis and treatment of hip joint pathologies) were selected as the source of prospective participants. Both sports physicians work in the primary health care sector, seeing patients as a 'first point of contact'. However, because of their recognised expertise with hip pathology, these physicians often see patients at a secondary care level whereby patients are referred to them by other health professionals (primarily physiotherapists and general practitioners). The orthopaedic surgeon works purely at the secondary care level, only seeing patients that are referred. This surgeon has specialised in hip arthroscopy over the last 15 years and has performed over 350 such procedures. Specialists were given full details about all aspects of this study. They agreed to ask consecutive patients who met the abovementioned primary inclusion criteria, if they would be interested in being included in this study and if so, for permission for their contact details to be given to the researcher.

### **5.5.5 Examiner**

One experienced examiner, a physiotherapist with 32 years experience and who is the author of this thesis, assessed all participants.

### **5.5.6 Procedures**

All patients referred by the medical specialists were telephoned by the researcher and given an overview of the study. Prospective participants were given the opportunity to have any questions relevant to the study answered and were then screened following a standardised questionnaire (see Appendix 10) to ensure that they met the inclusion/exclusion criteria detailed above. Those that met these criteria and who expressed a desire to be included in the study were emailed a number of documents including the detailed 'Study Information Sheet' (Appendix 11) and a 'Consent Form' (Appendix 12). Participants were asked to read the information sheet and to contact the researcher if they had any further questions or concerns about the study. Consent forms were not signed until the day of data collection, after the participant had been given another opportunity to discuss the study. The initial email also included a 'Medical Screening Questionnaire' (Appendix 13) designed to identify risk factors or medical conditions that would be a contraindication to participation in this study (including infection, cancer, deep vein thrombosis, fracture). Also attached to the email were validated questionnaires that provided information about activity levels (the Lower Limb Task Questionnaire) (McNair et al., 2007) (see Appendix 14), and the possibility of a neuropathic pain component to the participant's problem (The self-report version of the Leeds Assessment of Neuropathic Symptoms and Signs questionnaire [S-LANSS]) (Bennett et al., 2005; Weingarten et al., 2007) (see Appendix 15).

The researcher arranged an appointment time for the participant's MRA/FGAI and study data collection. The participant was instructed to avoid doing any physical activity 'out of the ordinary' and not to take any pain relief or anti-inflammatory medication in the 24 hours preceding data collection to minimise the effects that medication or unusual activity might have on their pain response during testing.

On the day of the patient's MRA/FGAI, experimental procedure was explained and informed consent acquired. Questionnaires were collected and checked for correct completion. Baseline data (including diagnosis, occupation, levels of activity, cause of current symptoms, aggravating and easing factors and any previous injury details) were collected via a standardised questionnaire (see Appendix 16). A body chart was used for

participants to indicate the site of pain (both primary and secondary) or associated symptoms (see Appendix 17). Baseline data collection was followed by a standardised clinical examination that included physical tests (see ‘Tests’ below) shown to be reliable in the preceding reliability studies. The initial examination was performed immediately prior to the FGAI and MRA. To minimise order effects, tests were performed in random order pre-determined by a random number generator (accessed at <http://www.pangloss.com/seidel/rnumber.cgi>).

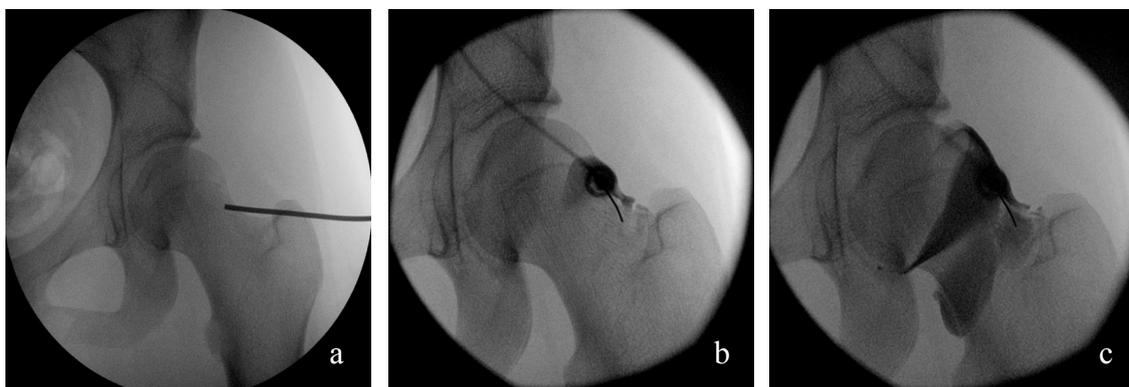
Participants were asked if the test provoked a ‘familiar’ pain i.e. pain that was very similar in nature and site to the pain that they usually felt at their hip. Participants rated the intensity of familiar pain using the 11-point numeric pain rating scale (NPRS). The NPRS was anchored at 0 (no pain) and 10 (worst pain imaginable). A positive test was defined as one that provoked familiar pain with an intensity of 2 or greater. Results for each test were recorded via an iPad (Apple Inc) utilising a standardised, interactive PDF form (see Appendix 18) created specifically for this study by the researcher (using PDF Expert (2008-2014) version 4.7.7.2; Readdle Inc). This ‘App’ allows data from numerical and text fields, checkboxes and radio buttons to be entered and saved within the form and converted to various file formats including MS Office Excel and to be exported directly to statistical software packages.

Immediately after the completion of the physical examination, the participant underwent a FGAI and then a MRA, in an adjacent room, following a standardised protocol (see FGAI and MRA Protocol below). The researcher re-examined the participant immediately after the MRA, repeating all tests in the same order as prior to the procedure. Again, the participants were required to report the reproduction of ‘familiar’ pain and to rate the intensity of such pain. The examiner was blinded to the results of the MRA at the time of reassessment and until after analysis of all data. The radiologist who performed the FGAI and MRA was blinded to the results of the clinical examination and anaesthetic response until after the completion of the study and all data analyses.

### **5.5.7 Reference test**

All patients underwent a FGAI and MRA performed by a musculoskeletal radiologist (with 30 years of experience) who was blinded to the results of the clinical examination. Intra-articular injection of anaesthetic was performed with aseptic technique under fluoroscopic guidance. The patient was placed in a supine position on the fluoroscopic

table with the x-ray tube positioned vertically under the table, the image intensifier above the patient and the affected hip and leg in a neutral position. The injection site (lateral to the femoral vessels, between the middle to superior part of the anatomical neck of the femur) was identified under fluoroscopic guidance using a radiopaque linear pointer (see Figure 5.1a). This site was marked on the skin with a marker pen. The skin was then cleaned with a suitable skin preparation (Chlohexidine gluconate or povidine-iodine solution) and a sterile field was obtained using a sterile window drape. A subcutaneous injection of local anaesthetic (Lidocaine 1%) was followed by the intra-articular placement of a 22-gauge spinal needle under fluoroscopic guidance (Figure 5.1b). Under fluoroscopic guidance, an intra-articular injection of the arthrogram mixture was initiated to adequately distend the joint capsule (Figure 5.1c).



**Figure 5.1 Fluoroscopy guided anaesthetic injection**

- a Identification of injection site by fluoroscopy
- b Injection of local anaesthetic
- c Injection of arthrogram mixture

The arthrogram mixture was a 1:200 concentration, comprised of 4ml iodinated contrast medium (Omnipaque™ 300, GE Healthcare, Iohexol 300mg Iodine per ml), 6ml bupivacaine hydrochloride 0.25% (Marcain, AstraZeneca Limited), 0.05ml chelated gadolinium (Magnevist® Bayer New Zealand Limited, 0.5 mmol/mL dimeglumine gadopentetate) and 0.3mg DBL™ Adrenalin 1:1000 (1mg in 1 ml, Hospira NZ Limited). MR imaging was performed as soon as possible after the intra-articular injection had been completed and the patient had been transported to the MR suite by wheelchair. Whilst all participants also underwent a MRA, the information from this procedure was used for secondary analysis of the data, not as the reference standard for the current study. Detail of the MRA procedure is provided in Chapter 7.

Pain response to the FGAI was used as the reference test. A positive response was considered to indicate that the participant's pain originated from an intra-articular pathology and a negative response considered to represent pain that originates from an extra-articular hip joint structure. A positive anaesthetic response (PAR) was determined, a priori, as an 80% or greater reduction in pain intensity of the mean score calculated from the individual scores of the three most provocative tests performed prior to the FGAI. Further detail in this regard is provided in the 'Data Analysis' section that follows.

### **5.5.8 Index tests**

Thirty individual tests were included in this study. Full descriptions of the included tests are provided in Appendix 7. Participants were also asked to demonstrate any functional activity or manoeuvre that they thought would be likely to reproduce their symptoms based on their day to day experience of their hip problem. If the pain was reproduced with such an activity, the patient was asked to repeat the activity a couple of times to see if the pain reproduction was consistent and then to score the intensity of the pain. This activity was called a 'Patient Specific Pain Provocation Manoeuvre' (PSPPM) and details were recorded along with those from the standardised tests.

### **5.5.9 Data analysis**

For each participant, a mean pain intensity score was calculated from the individual scores of the three most provocative tests performed prior to the MRA/FGAI. This score was compared to the mean score reported for the same three tests during the reassessment of the participant following the FGAI/MRA so that the response to the FGAI could be determined. Success from the injection (a positive anaesthetic response or PAR) was considered to be at least an 80% reduction in pain intensity of this mean score, provided that the change exceeded a two-point shift on the Numeric Pain Rating Scale (NPRS) (Farrar et al., 2001; Martin et al., 2008; Maslowski et al., 2010). Success from the injection was used as the reference standard for subsequent analyses.

Once all data collection was completed, 2 x 2 contingency tables were constructed using the Statistical Package for the Social Sciences (SPSS) software, version 22 (IBM© Corporation, 2013) to examine the diagnostic accuracy of individual tests. The dependent variable was the response to the anaesthetic (either positive or negative). Various measures of diagnostic accuracy were calculated including sensitivity, specificity, positive likelihood ratios (LR+) and negative likelihood ratios (LR-),

positive and negative predictive values and diagnostic odds ratios (DOR). 95% CIs were constructed for each variable using the Confidence Interval Calculator downloaded from the Physiotherapy Evidence Database (Herbert, 2013).

Continuous variables from the history and physical examination were examined for normality using the Kolmogorov-Smirnov test. All variables were examined for univariate relationship with the PAR using independent samples t-test for normally distributed continuous variables and Fishers Exact test for categorical variables. The Wilcoxon-Mann-Whitney test was used for continuous variables that were not normally distributed. Receiver operator characteristic (ROC) curves were constructed for continuous variables with a significant relationship to the PAR and sensitivity and specificity values were calculated for all possible cut-off points. Post-test probabilities were calculated for all variables.

## **5.6 Results**

Seventy-seven potential participants were referred for inclusion in this study. Two were excluded during the initial telephone-screening interview with one being too young (14 years) and the other being treated for concurrent low back pain with widespread pain referral. This left 75 participants who met the inclusion criteria, all of whom agreed to participate in the study. However, seven patients (4 males; age range 25-62 years) later withdrew. Three withdrew because they were claustrophobic and did not want to undergo the MRA procedure without sedation, a factor that would most likely influence their assessment of pain intensity. Three withdrew because their symptoms resolved prior to the agreed date for their data collection. One person withdrew because she did not think that her symptoms were severe enough to warrant further investigation. Six of the withdrawn patients reported that the onset of their hip pain was associated with trauma; the other could not identify the cause. Two of the withdrawn patients had a history of pain greater than 24 months, whereas the average time between initial injury and consultation with the referring specialist (duration) was 15 weeks for the remaining withdrawn patients.

Of the sixty-eight participants (32 males) included in the study, 25 were referred by the orthopaedic surgeon and 43 by sports physicians. Data on these individuals were collected between 28<sup>th</sup> June 2012 and 17<sup>th</sup> December 2013. Table 5.4 provides baseline information for all included participants, as well as a breakdown for the participants who had a positive response to the anaesthetic and those who did not. The mean age

across all participants was 38.2 years and mean body mass index 24.5. There was a statistically significant difference ( $p = 0.005$ ) in the mean age of the PAR group (42.1 years) and the NAR group (34.3 years). Thirty-four participants had a positive response ( $\geq 80\%$  reduction) to the anaesthetic injection and 34 had a negative response. Table 5.4 also provides detail regarding the mean pain intensity experienced by participants over the two days immediately prior to data collection. Best and worst pain intensity during the month before data collection is also provided. Twenty-seven of the participants that had a PAR reported at least a 90% reduction in pain intensity. Two participants had a mild (10-20%), temporary increase in pain after the anaesthetic (see Appendix 19 for more detail).

**Table 5.4 Baseline information**

Characteristic	All Cases (n=68)	PAR (n=34)	NAR (n=34)	p-value t-test
	Mean (SD)	Mean (SD)	Mean (SD)	
Age (years)	38.2 (11.8)	42.1(10.81)	34.3 (11.5)	0.005*
Height (metres)	1.74 (0.09)	1.73 (0.09)	1.74 (0.09)	0.79
Weight (kg)	74.9 (14.9)	76.3 (14.2)	73.4 (15.5)	0.43
BMI	24.5 (3.0)	25.1 (3.0)	23.9 (3.0)	0.11
Duration of symptoms (months)	21.0 (32.6)	21.0 (36.2)	20.0 (29)	0.87
Pain Intensity (Units on NPRS)				
Best over last month	0.3 (0.7)	0.4 (0.8)	0.4 (0.7)	0.94
Worst over last month	7.4 (2.0)	7.4 (2.3)	7.4 (1.7)	0.90
Average pain am last 2 days	1.9 (1.6)	1.9 (1.7)	2.0 (1.7)	0.83
Average pain pm last 2 days	2.6 (2.0)	2.5 (2.2)	2.6 (1.9)	0.88

PAR, Positive anaesthetic response; NAR, Negative anaesthetic response; SD, Standard Deviation; n, number of individuals; NPRS Numeric pain rating scale; \* Significant difference between PAR and NAR groups ( $p < 0.05$ )

Detail concerning participants referred by the orthopaedic surgeon is compared to those from the sports physicians in Table 5.5. There was not any statistically significant difference between groups for any of the variables reported in this table.

**Table 5.5 Comparison between surgeon and sports physician participants**

Characteristic	Surgeon (n=25)	Sports Physician (n=43)	p-value t-test
	Mean (SD)	Mean (SD)	
Age (years)	39.8 (10.3)	37.3 (12.6)	0.42
Height (metres)	1.75 (0.07)	1.72 (0.10)	0.18
Weight (kg)	77.1 (11.4)	73.5 (16.5)	0.35
BMI	24.8 (2.5)	24.3 (3.3)	0.53
Duration of symptoms (months)	17.0 (26)	23.6 (36)	0.42
Pain Intensity (units on NPRS)			
Best over last month	0.3 (0.6)	0.4 (0.8)	0.36
Worst over last month	7.2 (2.6)	7.5 (1.6)	0.57
Average pain am last 2 days	1.7 (1.6)	2.1 (1.7)	0.30
Average pain pm last 2 days	2.2 (2.2)	2.7 (2.0)	0.29
PAR Participants n (%)	14 (56)	20 (46)	0.61 <sup>1</sup>

SD, standard deviation; n, number of individuals; NPRS, numeric pain rating scale; <sup>1</sup> Fishers Test

Demographics of participants and detail from their history (relevant to hip pain) are reported in Table 5.6. Ninety percent of participants were New Zealand Europeans and the majority (84%) were employed at the time of data collection. Ninety-seven percent of participants reported that they normally participated in regular sport however 72% had stopped playing as a result of their current hip pain. None of these variables demonstrated a statistically significant relationship with a PAR.

**Table 5.6 Demographics and history**

	All Cases % <sup>1</sup> (n)	PAR Group % <sup>2</sup> (n)	NAR Group % <sup>2</sup> (n)	<i>p</i> -value Fishers Test
Ethnicity				
NZ European	90 (61)	85 (29)	94 (32)	0.43
Asian	4 (3)	3 (1)	6 (2)	1.0
Male gender	47 (32)	50 (17)	44 (15)	0.80
Employed	84 (57)	88 (30)	80 (27)	0.51
Left side painful	38 (26)	41 (14)	35 (12)	0.80
Left foot dominant	4 (3)	3 (1)	6 (2)	1.0
Plays regular sport	97 (66)	94 (32)	100 (34)	0.49
Stopped sport due to hip pain	72 (49)	61 (21)	82 (28)	0.10
Takes medication for hip regularly	16 (11)	21 (7)	12 (4)	0.51
Cause of current hip pain				
Trauma	62 (42)	62 (21)	62 (21)	1.0
Overuse	20 (14)	21 (7)	21 (7)	1.0
Unknown	18 (12)	18 (6)	18 (6)	1.0
History				
Previous problem same hip	37 (25)	41 (14)	32 (11)	0.61
Previous problem resolved	18 (12)	18 (6)	18 (6)	1.0
Previous problem other hip	25 (17)	29 (10)	21 (7)	0.57

<sup>1</sup> percent of total sample; <sup>2</sup> percent of subgroup; n, number of individuals

Table 5.7 provides detail in regard to activities that cause pain of greater than two points on the NPRS. The only activity that demonstrated a statistically significant difference ( $p = 0.017$ ) between groups was ‘pain when jogging’ which was more prevalent in the NAR group. This table also details the nature and prevalence of any associated symptoms. The most prevalent symptom was self-reported decreased ROM (reported by 75% of the whole cohort). The prevalence of this symptom was higher in the PAR group (88%) than the NAR group (62%) and this difference reached statistical significance ( $p = 0.023$ ). Similarly, the presence of crepitus was more prevalent in the PAR group (32%) than NAR group (9%) ( $p = 0.033$ ).

**Table 5.7 Aggravating activities & associated symptoms**

	All Cases % <sup>1</sup> (n)	PAR % <sup>2</sup> (n)	NAR % <sup>2</sup> (n)	<i>p</i> -value Fishers Test
Aggravating activities <sup>3</sup>				
Walking	56 (38)	50 (17)	62 (21)	0.46
Stairs	65 (44)	56 (19)	73 (25)	0.20
Standing	48 (33)	47 (16)	50 (17)	1.0
Sitting	37 (25)	35 (12)	38 (13)	1.0
Getting in/out of car	63 (43)	68 (23)	59 (20)	0.61
Putting on socks	50 (34)	47 (16)	53 (18)	0.80
Squatting	51 (35)	47 (16)	56 (19)	0.63
Driving	38 (26)	35 (12)	41 (14)	0.80
Jogging	78 (53)	65 (22)	91 (31)	0.017*
Twisting	79 (54)	82 (28)	76 (26)	0.76
Associated symptoms				
Crepitus	20 (14)	32 (11)	9 (3)	0.033*
Painful click	28 (19)	35 (12)	21 (7)	0.28
Painless click	44 (30)	50 (17)	38 (13)	0.46
Locking	19 (13)	15 (5)	23 (8)	0.54
Giving way/weakness	29 (20)	21 (7)	38 (13)	0.18
Self-reported ↓ROM	75 (51)	88 (30)	62 (21)	0.023*

<sup>1</sup> percent of total sample; <sup>2</sup> percent of subgroup; <sup>3</sup> activities that cause pain of 2 or greater on NPRS; ↓ROM, decreased range of movement; n, number of individuals \* Significant difference between PAR and NAR groups ( $p < 0.05$ )

The nature and area of pain is detailed in Table 5.8. The vast majority of participants (88%) reported intermittent pain (defined as ‘pain that was absent at some time during day’ as opposed to constant pain being ‘pain that never stops’). Similarly, the majority of participants (72%) reported that their dominant pain was felt in the groin. However, there was a statistically significant difference ( $p = 0.001$ ) between the PAR and NAR groups for this variable. Dominant groin pain was present in 91% of participants with a PAR compared to 53% of those with a NAR. There was not any statistically significant difference between groups for any of the other areas of pain distribution.

Forty percent ( $n=27$ ) of participants reported a secondary pain site that they considered was associated with their dominant pain, but which could be present with or without the dominant pain. The most common site of secondary pain for PAR participants was the mid-gluteal region (21%), whereas the groin was the most common secondary site for NAR participants (18%). Thirty-one percent of participants ( $n= 21$ ) reported referred pain that was present in association with their dominant pain. The most common site of referred pain was the anterior thigh. Of those that had referred pain, the majority (81%) had dominant pain in the groin. There was no statistically significant difference between PAR and NAR groups in regard to these variables.

**Table 5.8 Nature and area of pain**

	All Cases % <sup>1</sup> (n)	PAR % <sup>2</sup> (n)	NAR % <sup>2</sup> (n)	p-value Fishers Test
Description of pain (n=68)				
Intermittent pain	88 (60)	85 (29)	91 (31)	0.71
Dominant pain sharp	46 (31)	59 (20)	32 (11)	0.51
Dominant pain ache	59 (40)	50 (17)	68 (23)	0.22
Dominant pain region (n=68)				
Groin	72 (49)	91 (31)	53 (18)*	0.001*
TFL	13 (9)	9 (3)	18 (6)	0.48
Gluteal	9 (6)	6 (2)	12 (4)	0.67
Trochanteric	9 (6)	3 (1)	15 (5)	0.19
Secondary pain region (n=27)				
Groin	10 (7)	3 (1)	18 (6)	0.10
Upper gluteal	3 (2)	6 (2)	0 (0)	0.49
Mid-gluteal	15 (10)	21 (7)	9 (3)	0.30
Ischial	6 (4)	6 (2)	6 (2)	1.0
SIJ	4 (3)	9 (3)	0 (0)	0.24
Anterior thigh	1.5 (1)	0 (0)	3 (1)	1.0
Trochanteric	4 (3)	3 (1)	6 (2)	1.0
Referred pain (n=21)				
Anterior thigh	13 (9)	15 (5)	12 (4)	1.0
Posterior thigh	1.5 (1)	0 (0)	3 (1)	1.0
ITB	10 (7)	9 (3)	12 (4)	1.0
Adductor	4 (3)	0 (0)	9 (3)	0.24
Anterior knee	4 (3)	3 (1)	6 (2)	1.0
Testicle	6 (4)	9 (3)	3 (1)	0.61

<sup>1</sup> percent of total sample; <sup>2</sup> percent of subgroup;

n, number of individuals \* Significant difference between PAR and NAR groups (p < 0.05)

The functional status of the participants is reported in Table 5.9. Mean scores for the lower limb task questionnaire (LLTQ) demonstrate that recreational activities (RA) were more restricted than activities of daily living (ADL) for the whole cohort as well as for both the PAR and NAR groups. No participant scored less than 15/40 on the ADL and only 5 scored 20 or below. This contrasts with the RA sub-score for which 22 participants scored 20 or below. Scores on the Self-completed Leeds Assessment of Neuropathic Symptoms and Signs (S-LANSS) questionnaire were generally very low, with a mean of just three out of a possible maximum of 24. Thirty-seven participants scored zero, twenty eight scored between 3 and 11 and only three participants scored 12 or higher, the recommended cut-off point for the identification of neuropathic pain. All three participants who scored above this cut-off point had negative responses to the anaesthetic (percent change in pain intensity ranging from 28.5 to 35.1). There was no statistical difference in the LLTQ or S-LANSS scores between PAR and NAR subgroups or between the participants referred by the orthopaedic surgeon and those referred from the sports physicians.

**Table 5.9 Functional and neuropathic pain status**

Questionnaire	All Cases (n=68)	PAR (n=34)	NAR (n=34)	Surgeon (n=25)	Physician (n=43)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Functional Status (LLTQ)					
ADL score <sup>1</sup>	32.9 (6.5)	32.5 (7.1)	33.3 (6.0)	32.9 (7.0)	32.8 (6.4)
Recreational score <sup>1</sup>	23.8 (8.9)	23.2 (9.6)	24.3 (8.2)	23 (9.5)	24.2 (8.6)
Total score <sup>2</sup>	56.6 (14.3)	55.7 (15.7)	57.6 (13)	56 (15.9)	57 (13.6)
Neuropathic Pain Status					
S-LANSS Score <sup>3</sup>	3 (4.5)	2 (3.4)	4 (5.3)	3.3 (4.4)	2.8 (4.6)

LLTQ, Lower Limb Task Questionnaire; S-LANSS, Self-completed Leeds assessment of neuropathic symptoms and signs

<sup>1</sup> Maximum score = 40; <sup>2</sup> Maximum score = 80; <sup>3</sup> Maximum score = 24

Detail regarding mean ROM for both the PAR and NAR groups is displayed in Table 5.10. In respect to the painful hip, prior to the FGAI, there was a statistically significant difference between groups for mean ROM of internal rotation at 90° flexion ( $p = 0.024$ ) with the PAR group demonstrating less range (34°; SD 9.6°) than the NAR group (40°; SD 11.1°). Table 5.10 also provides detail regarding the difference in mean ROM between the painful and the non-painful hips prior to the FGAI. A positive value indicates that the non-painful side had greater ROM than the painful hip side. For both the PAR and NAR groups, the non-painful hip had more ROM than the painful hip, except for IR at 90° where the painful hip had slightly more ROM. When comparing these differences between sides across PAR and NAR groups, they were statistically significant for both the BKFO (5.5° versus 1.3°;  $p = 0.02$ ) and internal rotation (3.9° versus -0.6°;  $p = 0.01$ ).

Following the FGAI, mean ROM increased across all movements (except for external rotation in the PAR group) in both the PAR and NAR participants (see Table 5.10). These increases in ROM were larger in PAR participants and were statistically significant for both the BKFO (5.7° increase for PAR participants versus 2.5° increase for NAR;  $p = 0.03$ ) and internal rotation at 90° flexion (3.0° versus 0.2°;  $p = 0.01$ ).

**Table 5.10 Range of movement in degrees**

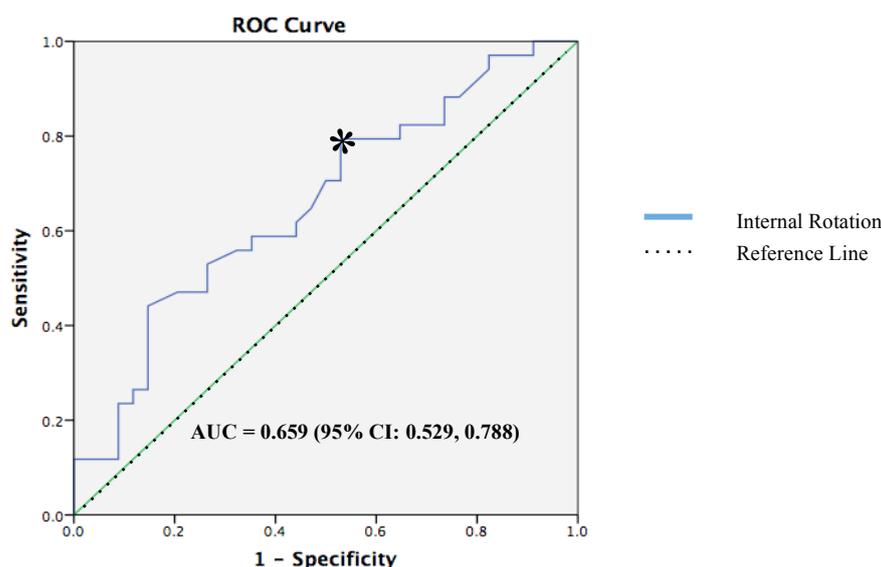
	<b>PAR (n=34)</b>	<b>NAR (n=34)</b>	<b>p-value</b>
	<b>Mean (SD)</b>	<b>Mean (SD)</b>	<b>t-test</b>
<b>Mean ROM of painful hip pre-FGAI</b>			
BKFO	58 (9.6)	62 (6.7)	0.06
Flexion	110 (7.0)	112 (9.5)	0.35
Internal rotation @ 90 <sup>0</sup> flexion	34 (9.6)	40 (11.1)	0.02*
External rotation @ 90 <sup>0</sup> flexion	39 (6.7)	42 (6.0)	0.14
<b>Difference in ROM between painful and non-painful hips pre-FGAI<sup>1</sup></b>			
BKFO	5.5 (7.6)	1.3 (6.2)	0.02*
Flexion	3.3 (7.5)	1.8 (5.6)	0.37
Internal rotation @ 90 <sup>0</sup> flexion	3.9 (6.9)	-0.6 (6.2)	0.01*
External rotation @ 90 <sup>0</sup> flexion	1.9 (5.1)	1.7 (5.4)	0.90
<b>Change in ROM of painful hip post-FGAI<sup>2</sup></b>			
BKFO	5.7 (6.9)	2.5 (4.5)	0.03*
Flexion	3.3 (5.6)	0.8 (5.1)	0.07
Internal rotation @ 90 <sup>0</sup> flexion	3.0 (4.3)	0.2 (3.9)	0.01*
External rotation @ 90 <sup>0</sup> flexion	-0.3 (4.1)	0.2 (2.5)	0.60

\* Significant difference between PAR and NAR groups ( $p < 0.05$ ); BKFO, bent knee fall out

<sup>1</sup> Positive result indicates that non-painful hip has more ROM than painful hip

<sup>2</sup> Positive result indicates that ROM increased post-FGAI

Receiver operating characteristic (ROC) curves were generated for continuous variables that demonstrated a statistically significant association with the PAR. Figure 5.2 shows the ROC curve for mean internal rotation ROM. This analysis identified that the best combination of sensitivity and specificity for internal rotation ROM was when the range was 40.8<sup>0</sup> (indicated by the \*). Sensitivity at this point was 0.79 and specificity 0.44. The area under the ROC curve (AUROC) for internal rotation ROM was 0.66 (95% CI 0.53, 0.79). The post-test probability of a PAR at this point increased from 50% to 59% (see Table 5.17). Specificity increased to a maximum of 1.0 when internal ROM was  $\leq 20^0$ , although sensitivity at this value dropped to just 0.12 (see Appendix 20).

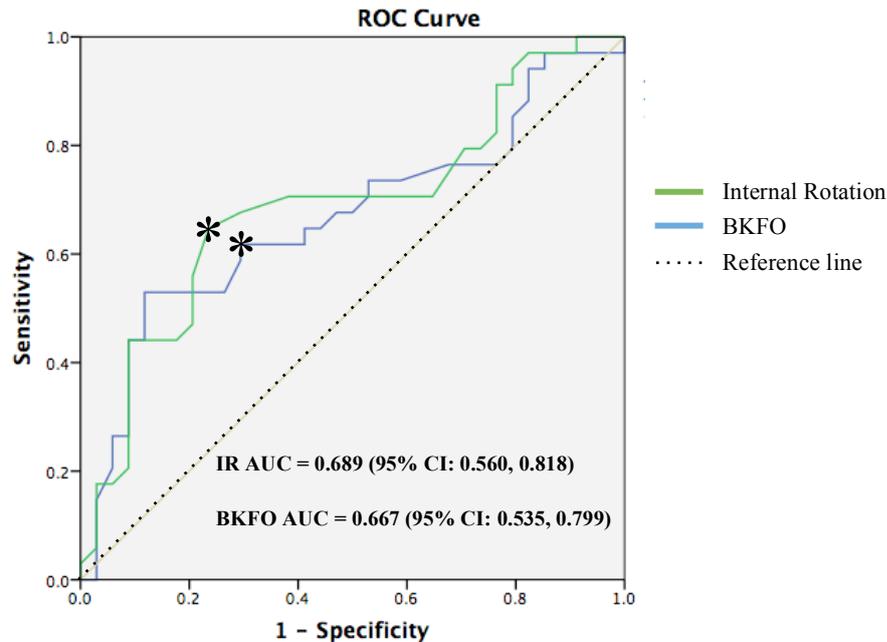


**Figure 5.2 ROC curve for mean range of internal rotation of the painful hip**

\* Indicates the point of best overall accuracy for this variable.  
Abbreviations: AUC, area under curve; CI, confidence intervals

Figure 5.3 shows the ROC curve for the *pre-FGAI differences* in ROM between the painful and non-painful hips for internal rotation. The best combination of sensitivity and specificity for internal rotation was when the painful hip had  $2^{\circ}$  of movement less than the non-painful hip (sensitivity 0.65, specificity 0.76, LR+ 2.7, LR- 0.46). The post-test probability of a PAR at this point increased from 50% to 73% (see Table 5.17). Specificity increased as the size of this difference increased, reaching 1.0 when the painful hip had  $16^{\circ}$  less range than the non-painful hip. Sensitivity at this point was only 0.03 (Appendix 21).

Figure 5.3 also shows the ROC curve for the *pre-FGAI differences* in ROM between the painful and non-painful hips for the BKFO test. The best combination of sensitivity and specificity for this movement was when the painful hip had  $4^{\circ}$  of movement less than the non-painful hip (sensitivity 0.62, specificity 0.70, LR+ 2.0, LR- 0.54). The post-test probability of a PAR at this point increased from 50% to 67% (see Table 5.17). Specificity increased as the size of this difference increased, reaching 1.0 when the painful hip had  $20^{\circ}$  less range than the non-painful hip. However, sensitivity at this point was zero (see Appendix 21).



**Figure 5.3 ROC curve for mean *differences* in range of movement between painful and non-painful hips prior to the FGAI**

\* Indicates the point of best overall accuracy for each variable.  
Abbreviations: AUC, area under curve; CI, confidence intervals;

With respect to participant age, the best combination of sensitivity and specificity was when the patient was at age 39 years (ROC curve not shown). At this cut-off point, sensitivity was 0.70, specificity was 0.65 (see Appendix 22), the LR+ was 2.0 and the LR- was 0.45. The AUROC for age was 0.70 (95% CI: 0.58, 0.83). Post-test probability of a PAR at this point increased to 67% (see Table 5.17). Specificity increased to a maximum of 1.0 at age 60 years. Sensitivity at this point was only 0.03.

ROC analysis was also performed for mean ROM of the BKFO test considering how close this variable came to having a statistical relationship with the PAR. This analysis identified that the best combination of sensitivity and specificity for this test was when the range was  $\leq 62^{\circ}$ . At this cut-off point, sensitivity was 0.67, specificity was 0.56 (see Appendix 23) the LR+ was 1.5 and the LR- was 0.59. The AUROC for this variable was 0.67 (95% CI: 0.53, 0.80). Post-test probability of a PAR at this point increased to 60% (see Table 5.17). Specificity increased to a maximum of 1.0 when BKFO ROM was  $\leq 47^{\circ}$ , although sensitivity at this value dropped to just 0.09.

The following tables provide detail about the diagnostic accuracy of the physical tests included in the current study. Table 5.11 summarises values for the ‘Impingement Test’ and some of the variations of this test described in the literature. These tests demonstrate high sensitivity, ranging from 0.79 (95% CI 0.63, 0.90) for compression through the long axis of the femur with the hip at 90<sup>0</sup> flexion and slight adduction (FADC) to 0.97 (95% CI 0.85, 0.99) for the quadrant test. The high sensitivity of the quadrant test along with the low negative LR (0.14) indicates that a negative response for this test might be useful for ruling out a PAR. However, the wide confidence intervals for this LR mean that this may not be the case. This test just failed to demonstrate a statistically significant association with a PAR with a *p*-value of 0.054 determined by Fishers Exact test. In contrast to the high sensitivity, specificity for the ‘impingement tests’ is low, ranging from 0.18 (95% CI 0.08, 0.34) for the FADDIR test to 0.35 (95% CI 0.21, 0.52) for FADC.

The resisted tests (as pain provocation tests as opposed to measures of strength) generally demonstrate *unacceptable* diagnostic accuracy values (see Table 5.12). The specificity of resisted extension suggests that it might be useful clinically to help rule in a PAR. However, the very low value of the associated positive LR does not support this suggestion. Two of these tests, resisted abduction and the external derotation test (EDT), demonstrated a statistically significant relationship with a PAR. However, none of the metrics used to indicate the performance of these tests indicate that they have diagnostic utility.

Table 5.13 summarises the accuracy values for tests that explore the effect of end range rotation (internal and external) in various degrees of hip flexion. Internal rotation in full flexion (FFIR) has the highest diagnostic utility out of this group of tests with a sensitivity of 0.91 (95% CI 0.77, 0.97), a negative LR of 0.33 (95% CI 0.10, 1.13) and a diagnostic odds ratio of 3.72 (95% CI 0.91, 15.22). The sensitivity of internal rotation at 90<sup>0</sup> flexion (FIR) and the specificity of external rotation in sitting (0.79 and 0.74 respectively) indicate that these tests may have some diagnostic value. Unfortunately, poor LR values indicate that neither of these tests is likely to be of great value clinically. Tests performed in ‘weight-bearing’ demonstrate poor diagnostic accuracy values (see Table 5.14). All positive LR’s and DOR’s are less than 1 and all negative LR’s are greater than 1. Similarly, sensitivity and specificity are too low to suggest that these tests are useful for identifying or ruling out intra-articular pathology of the hip. Of the remaining physical tests (detailed in Table 5.15), the log roll test has the highest

diagnostic accuracy values with a specificity of 0.85 (95% CI 0.70, 0.94), a positive LR of 1.6 (95% CI 0.58, 4.4) and DOR of 1.8 (95% CI 0.52, 6.15). Specificity of passive extension in prone was high (0.82 (95% CI 0.66, 0.92)) and the sensitivity the BKFO test was reasonable (0.71 (95% CI 0.54, 0.83)).

Five variables collected from the participants 'history' demonstrated a statistically significant relationship with a PAR ( $p < 0.05$ ). Diagnostic accuracy values were calculated for these variables and are presented in Table 5.16. Sensitivity was high for 'dominant pain in the groin' (0.91 (95% CI 0.77, 0.97)) and 'self reported limitation of movement' (0.85 (95% CI 0.70, 0.94)). Specificity was high for the 'presence of crepitus' (0.91 (95% CI 0.77, 0.97)). The presence of crepitus had the highest positive LR (3.67) and dominant pain in the groin had the best negative LR (0.18 (95% CI 0.06, 0.58)) and diagnostic odds ratio (9.18).

**Table 5.11 Diagnostic accuracy of ‘impingement’ tests**

Test	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	p-value
FADDIR	31	28	3	6	0.91 (0.77, 0.97)	0.18 (0.08, 0.34)	0.53/0.67	1.11 (0.92, 1.34)	0.5 (0.14, 1.84)	2.21 (0.51, 9.70)	0.476
Quadrant	33	27	1	7	0.97 (0.85, 0.99)	0.21 (0.10, 0.37)	0.55/0.88	1.22 (1.02, 1.47)	0.14 (0.02, 1.10)	8.56 (0.99, 73.9)	0.054
FADC	27	22	7	12	0.79 (0.63, 0.90)	0.35 (0.21, 0.52)	0.55/0.63	1.23 (0.91, 1.66)	0.58 (0.26, 1.30)	2.1 (0.71, 6.25)	0.280
FF	29	23	5	11	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	0.56/0.69	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)	0.152

TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio.

FADDIR, 90° Flexion Adduction Internal Rotation; FF, Full Flexion; FADC, Flexion Adduction Compression

**Table 5.12 Diagnostic accuracy of resisted movements (as pain provocation tests)**

Test	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	p-value
RAD	7	15	27	19	0.21 (0.10, 0.37)	0.56 (0.39, 0.71)	0.32/0.41	0.47 (0.22, 1.0)	1.42 (1.01, 2.01)	0.33 (0.11, 0.96)	0.068
RAB	1	8	33	26	0.03 (0.01, 0.15)	0.76 (0.60, 0.88)	0.11/0.44	0.13 (0.02, 0.95)	1.27 (1.04, 1.54)	0.10 (0.01, 0.84)	0.043*
RF	5	11	29	23	0.15 (0.06, 0.30)	0.68 (0.51, 0.81)	0.31/0.44	0.45 (0.18, 1.17)	1.26 (0.96, 1.65)	0.36 (0.11, 1.19)	0.152
RE	4	5	30	29	0.12 (0.05, 0.27)	0.85 (0.70, 0.94)	0.44/0.49	0.80 (0.24, 2.73)	1.03 (0.86, 1.25)	0.77 (0.19, 3.17)	1.00
RIR	6	11	28	23	0.18 (0.08, 0.34)	0.68 (0.51, 0.81)	0.35/0.45	0.55 (0.23, 1.31)	1.22 (0.92, 1.61)	0.45 (0.14, 1.4)	0.262
RER	5	10	29	24	0.15 (0.06, 0.30)	0.71 (0.54, 0.83)	0.33/0.45	0.50 (0.19, 1.31)	1.21 (0.93, 1.56)	0.41 (0.12, 1.38)	0.242
EDT	2	11	32	23	0.06 (0.02, 0.19)	0.68 (0.51, 0.81)	0.15/0.42	0.18 (0.04, 0.76)	1.39 (1.09, 1.78)	0.13 (0.03, 0.65)	0.011*

TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio; \* Statistically significant association with PAR ( $p < 0.05$ ) calculated by Fishers Exact test;

RAD, Resisted Adduction; RAB, Resisted Abduction; RF, Resisted Flexion; RE, Resisted Extension; RIR, Resisted Internal Rotation @ 90°; RER, Resisted External Rotation @ 90°; EDT, External Derotation Test

**Table 5.13 Diagnostic accuracy of end-range ‘rotation’ tests**

Test	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	<i>p</i> -value
<b>Internal Rotation Tests</b>											
FFIR	31	25	3	9	0.91 (0.77, 0.97)	0.26 (0.15, 0.43)	0.55/0.75	1.24 (0.99, 1.56)	0.33 (0.10, 1.13)	3.72 (0.91, 15.22)	0.109
FIR	27	28	7	6	0.79 (0.63, 0.90)	0.18 (0.08, 0.34)	0.49/0.46	0.96 (0.77, 1.22)	1.17 (0.44, 3.11)	0.8 (0.25, 2.78)	1.00
IRSit	16	17	18	17	0.47 (0.31, 0.63)	0.50 (0.34, 0.66)	0.48/0.49	0.94 (0.58, 1.54)	1.06 (0.67, 1.68)	0.89 (0.34, 2.30)	1.00
IRP	17	19	17	15	0.50 (0.34, 0.66)	0.44 (0.29, 0.61)	0.47/0.47	0.89 (0.57, 1.40)	1.13 (0.68, 1.88)	0.79 (0.30, 2.05)	0.808
IRSt	12	19	22	15	0.35 (0.21, 0.52)	0.44 (0.29, 0.61)	0.39/0.41	0.63 (0.37, 1.09)	1.47 (0.93, 2.31)	0.43 (0.16, 1.14)	0.144
<b>External Rotation Tests</b>											
FFER	15	14	19	20	0.44 (0.29, 0.61)	0.59 (0.42, 0.74)	0.52/0.51	1.07 (0.62, 1.86)	0.95 (0.63, 1.43)	1.13 (0.43, 2.95)	1.00
FER	16	14	18	20	0.47 (0.31, 0.63)	0.59 (0.42, 0.74)	0.53/0.53	1.14 (0.67, 1.96)	0.9 (0.59, 1.37)	1.27 (0.49, 3.3)	0.807
ERSit	12	9	22	25	0.35 (0.21, 0.52)	0.74 (0.57, 0.85)	0.57/0.53	1.33 (0.65, 2.74)	0.88 (0.64, 1.21)	1.52 (0.54, 4.27)	0.410
ERP	10	13	24	21	0.29 (0.17, 0.46)	0.62 (0.45, 0.76)	0.43/0.47	0.77 (0.39, 1.51)	1.14 (0.81, 1.61)	0.67 (0.25, 1.85)	0.609
ERSt	7	11	27	23	0.21 (0.10, 0.37)	0.68 (0.51, 0.81)	0.39/0.46	0.64 (0.28, 1.45)	1.17 (0.88, 1.57)	0.54 (0.18, 1.63)	0.600

TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio.

FFIR, Full Flexion Internal Rotation; FIR, Flexion Internal Rotation at 90° Flexion; IRSit, Internal Rotation in Sitting; IRP, Internal Rotation in Prone; IRSt Internal Rotation in Standing;

FFER, Full Flexion External Rotation; FER, Flexion External Rotation at 90° Flexion; ERSit, External Rotation in Sitting; ERP, External Rotation in Prone; ERSt External Rotation in Standing.

**Table 5.14 Diagnostic accuracy of ‘weight-bearing’ tests**

Test	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	<i>p</i> -value
ADDSt	18	20	16	14	0.53 (0.37, 0.69)	0.41 (0.26, 0.58)	0.47/0.47	0.90 (0.59, 1.38)	1.14 (0.67, 1.96)	0.79 (0.30, 2.06)	0.807
SOLSt	6	13	28	21	0.18 (0.08, 0.34)	0.62 (0.45, 0.76)	0.32/0.43	0.46 (0.20, 1.07)	1.33 (0.98, 1.81)	0.35 (0.11, 1.06)	0.104
4PtFlex	10	15	24	19	0.29 (0.17, 0.46)	0.56 (0.39, 0.71)	0.40/0.44	0.67 (0.35, 1.27)	1.26 (0.87, 1.83)	0.53 (0.19, 1.44)	0.314

TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio.

ADDSt, Adduction in Standing; SOLSt, Sustained One Leg Standing, 4PtFlex, Hip flexion in 4 point kneel.

**Table 5.15 Diagnostic accuracy of ‘miscellaneous’ tests**

Test	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	<i>p</i> -value
FABER	23	24	11	10	0.68 (0.51, 0.81)	0.29 (0.17, 0.46)	0.49/0.48	0.96 (0.70, 1.32)	1.10 (0.54, 2.24)	0.87 (0.31, 2.44)	1.00
BKFO	24	23	10	11	0.71 (0.54, 0.83)	0.32 (0.19, 0.49)	0.51/0.52	1.04 (0.76, 1.43)	0.91 (0.45, 1.85)	1.15 (0.41, 3.21)	1.00
PExt	7	6	27	28	0.21 (0.10, 0.37)	0.82 (0.66, 0.92)	0.54/0.51	1.17 (0.44, 3.11)	0.96 (0.77, 1.22)	1.21 (0.36, 4.07)	1.00
Log Roll	8	5	26	29	0.24 (0.12, 0.40)	0.85 (0.70, 0.94)	0.62/0.53	1.60 (0.58, 4.4)	0.90 (0.71, 1.13)	1.80 (0.52, 6.15)	0.539
TOP GT	6	12	28	22	0.18 (0.08, 0.34)	0.65 (0.48, 0.79)	0.33/0.44	0.50 (0.21, 1.18)	1.27 (0.95, 1.71)	0.39 (0.13, 1.21)	0.168

TP True Positives; FP False Positives; FN False Negatives, TN True Negatives; CI Confidence Intervals; PPV Positive Predictive Value; NPV Negative Predictive Value; LR+ Positive Likelihood Ratio; LR- Negative Likelihood Ratio; DOR Diagnostic Odds Ratio  
 FABER, Flexion Abduction External Rotation; BKFO, Bent Knee Fall Out; PExt, Passive Extension Prone; TOP GT, Tenderness on Palpation of Greater Trochanter

**Table 5.16 Diagnostic accuracy of variables from history that have a significant association with PAR ( $p < 0.05$ )**

Variable	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	PPV/NPV	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)	<i>p</i> -value
Crepitus	11	3	23	31	0.32 (0.19, 0.49)	0.91 (0.77, 0.97)	0.79/0.57	3.67 (1.12, 11.99)	0.74 (0.58, 0.96)	4.94 (1.24, 19.76)	0.033*
Age $\geq$ 39	24	12	10	22	0.70 (0.54, 0.83)	0.65 (0.48, 0.79)	0.67/0.69	2.00 (1.20, 3.31)	0.45 (0.26, 0.81)	4.40 (1.59, 12.19)	0.015*
SR $\downarrow$ ROM	29	23	5	11	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	0.56/0.69	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)	0.023*
Pain jogging <sup>1</sup>	22	31	12	3	0.65 (0.48, 0.79)	0.09 (0.03, 0.23)	0.42/0.20	0.71 (0.54, 0.93)	4.0 (1.24, 12.92)	0.18 (0.05, 0.70)	0.017*
Dominant pain in groin	31	18	3	16	0.91 (0.77, 0.97)	0.47 (0.31, 0.63)	0.63/0.84	1.72 (1.23, 2.04)	0.18 (0.06, 0.58)	9.18 (2.35, 35.89)	0.001*

TP, True Positives; FP, False Positives; FN, False Negatives, TN, True Negatives; CI, Confidence Intervals; PPV, Positive Predictive Value; NPV, Negative Predictive Value; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio; \* Statistically significant association with PAR ( $p < 0.05$ ) calculated by Fishers Exact test; <sup>1</sup> Pain of 2 or greater on NPRS; SR $\downarrow$ ROM, Self-reported decreased range of movement

Table 5.17 shows the changes in probability of a PAR, given a positive or negative test, for variables that had a DOR of greater than 1.8. Also included are the ROM variables that had a statistically significant association with a PAR, using the cut-off points identified by the ROC analysis. Pre-test probability of 50% is based on the number of participants in the study who had a PAR (34 out of 68). Post-test probability for positive test results ranged from 53% (for a positive FADDIR test) to 79% (for the presence of crepitus). Given a negative test result, the probability of a PAR dropped to as low as 12% (for a negative Quadrant test) and 16% (for not having dominant pain in the groin).

**Table 5.17 Post-test probability of a positive anaesthetic response**

Variable	Positive Test Result Post-Test Probability (%)	Negative Test Result Post-Test Probability (%)
FADDIR	53	33
Quadrant test	55	12
FADC	55	37
FFIR	55	26
FF	56	32
Patient reported limited movement	56	32
IR90 ROM < 41 <sup>0</sup>	59	32
BKFO < 62 <sup>0</sup>	60	37
Log roll	62	47
Dominant pain in groin	63	16
Age ≥ 39	67	31
<sup>1</sup> Difference in BKFO ROM between sides ≥ 4 <sup>0</sup>	67	35
<sup>1</sup> Difference in IR ROM between sides ≥ 2 <sup>0</sup>	73	32
Crepitus	79	43

Pre-test Probability = 50% (Prevalence of PAR in this study)

<sup>1</sup> Where pre-FGAI ROM of the painful hip is less than non-painful hip ROM

FADDIR, Flexion Adduction Internal Rotation; FADC, Flexion Adduction Compression; FFIR, Full Flexion Internal Rotation; FF, Full Flexion; IR90 ROM, Range of Movement of Internal Rotation with hip flexed 90<sup>0</sup>; BKFO, Bent Knee Fall Out; IR, Internal Rotation; ROM, Range of Movement

## 5.7 Discussion

This study provides new evidence that will improve our ability to diagnose intra-articular hip joint pathology. Firstly, it provides information regarding the pre-test probability of a positive anaesthetic response (PAR) and by implication, the presence of intra-articular pathology of the hip. Overall, 50% of participants in the current study had a reduction in pain of  $\geq 80\%$  following the injection. We can compare this finding to two previous diagnostic accuracy studies (Martin et al., 2008; Maslowski et al., 2010) of the hip that have used an intra-articular injection of anaesthetic as the reference standard. Consistent with the current study, Maslowski and colleagues defined a PAR as an 80% reduction in pain intensity following this procedure. These authors reported that 40% of their participants had a PAR. Martin et al. defined a PAR as a  $>50\%$  reduction in pain and reported a slightly higher prevalence of 55%. These authors also reported data that demonstrated that 15% of participants had  $>90\%$  reduction in pain intensity and 45% had  $\geq 75\%$  reduction in pain intensity. Not surprisingly, this data demonstrates that the prevalence of a PAR increases as the level of pain intensity used to define such a response decreases. Setting this bar too low will result in a much larger number of patients being identified as having intra-articular pathology. A consequence of this may be that more patients will undergo unnecessary expensive and invasive diagnostic procedures such as MRA and surgical exploration of their hip. Our strict requirement of  $\geq 80\%$  reduction in pain is consistent with international guidelines (Bogduk, 2004b) for diagnostic injections and we believe that the prevalence that we have reported provides an accurate indication of the likelihood of the presence of symptomatic intra-articular pathology.

The prevalence of a PAR was not statistically different in the participants referred by the orthopaedic surgeon compared to those referred by the sports physicians. Similarly, a comparison of group means show that there were no statistical differences in the duration that symptoms had been present, the pain intensity or the functional status between these two groups. Whilst sports physicians and orthopaedic surgeons can assume that the prevalence of an intra-articular source of hip pain in the patients they see with hip pain is around 50%, general practitioners and physiotherapists should take into consideration that the prevalence of such pathology in a purely primary care practice may differ from the prevalence seen in this study.

This study also extends our knowledge regarding ROM of painful hips. Considering mean range, PAR participants demonstrated less internal rotation at  $90^{\circ}$  of flexion than

NAR participants (mean  $34^{\circ}$  and  $40^{\circ}$  respectively;  $p = 0.024$ ). Also, the increase in range for PAR participants after injection of anaesthetic into the hip joint was significantly ( $p = 0.01$ ) greater than that for NAR participants for both internal rotation at  $90^{\circ}$  (PAR group  $3.0^{\circ}$ ; NAR group  $0.2^{\circ}$ ) and the BKFO tests (PAR group  $5.7^{\circ}$ ; NAR group  $2.5^{\circ}$ ). The increases in range following this procedure suggest that the restriction in movement was at least in part due to the effect of pain. Indeed, the mean improvement in range with each movement tested in the PAR participants was very similar in degree to the differences in movement observed between hips prior to the FGAI. The relative *lack* of increase in range with the NAR participants following the FGAI may indicate that there may be shortening of the hip muscles or capsule, or perhaps a bony block to movement in this group. Alternatively, the lesser degree of pain relief that NAR participants experienced may not have been sufficient to modify movement restriction secondary to pain.

This relationship between a loss of internal rotation of the hip and intra-articular pathology of the hip is consistent with the findings of other studies. However, the degree of limitation considered important across these studies ranges from 15 to 28 degrees (Altman et al., 1991; Birrell et al., 2001; Chong et al., 2013; Holla et al., 2012; Sutlive et al., 2008). The findings of Altman et al. (1991) formed the basis for the American College of Rheumatology (ACR) criteria for the classification of patients with hip pain associated with OA. These criteria utilise  $< 15^{\circ}$  of internal rotation as a cut-off point in their classification tree although it is not clear how this degree of range was identified. Holla et al. (2012) used ROC curves to identify the cut-off point with the highest discriminative ability based on maximising both sensitivity and specificity. They reported that when hip IR was  $< 24^{\circ}$ , the probability of osteophytes or joint space narrowing (as evidence for OA) increased from 25% to 46% and decreased to 16% when internal rotation was  $\geq 24^{\circ}$ . Holla and colleagues also calculated post-test probabilities using the ACR cut-point of  $15^{\circ}$ . They reported that the probability of radiological evidence of OA being identified increased to 58% when internal rotation was  $< 15^{\circ}$  and decreased to 22% when  $\geq 15^{\circ}$ .

We employed the same analysis as Holla et al. to determine the cut-off point for range of internal rotation that *optimised* both sensitivity and specificity. Our analysis determined a cut-off point of  $40.8^{\circ}$ . The probability of a PAR increased from 50% (the pre-test prevalence) to 59% when internal rotation was  $< 41^{\circ}$  and decreased to 32% when it was greater than this figure. To provide further comparison to Holla et al., a cut-

point of  $< 24^{\circ}$  in our study increased the post-test probability of a PAR from 50% to 57%. Similarly, the post-test probability decreased to 44% when internal rotation was  $> 24^{\circ}$ . Whilst these comparisons indicate that there is inconsistency in the actual degree of limitation of internal rotation ROM that has the greatest diagnostic utility, there is clearly a consensus that this impairment is associated with hip pathology. The differences in the identified cut-points are most likely a result of methodological differences across studies (including the position and method of measuring internal rotation), differences in the characteristics of the participants (age range, diagnoses) and the various reference standards used (clinical diagnosis of OA, radiological diagnosis of OA, FGAI).

The current study is the first study to consider the diagnostic utility of differences in ROM between the painful and non-painful sides. We demonstrated that when the painful hip has  $4^{\circ}$  less range than the non-painful hip for the BKFO test, or  $2^{\circ}$  less range of internal rotation (tested at  $90^{\circ}$  flexion) the post-test probability of a PAR increases to just around 70%. However, given the SEM of  $1.6^{\circ}$  and  $2.1^{\circ}$  (respectively) associated with within-session measurements of ROM that we demonstrated in Chapter 4 (see Table 4.3), this finding may not have clinical utility.

Of the tests included in the physical examination, the quadrant test appears to have the highest diagnostic utility with a DOR of 8.56, a high sensitivity of 0.97 (95% CI 0.85, 0.99) and a low negative likelihood ratio (LR) of 0.14 (95% CI 0.02, 1.10). A negative result for this test reduced the probability of a PAR from 50% to 12%. This suggests that this test is useful for helping to 'rule out' the presence of intra-articular pathology, although the upper confidence interval for the negative LR is too high to be absolutely confident that this is the case. The specificity of 0.21 (95% CI 0.10, 0.37) and positive LR of 1.22 (95% CI 1.02, 1.47) for this test are too low for it to be considered useful for identifying such pathology as a stand-alone test. Our results for this test can be compared to those of Maslowski et al. (2010) who used the same reference standard as we employed ( $\geq 80\%$  reduction in pain following FGAI). These authors reported sensitivity of 0.50 (95% CI 0.26, 0.74) and specificity of 0.29 (95% CI 0.12, 0.51). Whilst this level of specificity is very similar to our finding, the sensitivity is much lower. There are a number of factors that might explain this difference. The most likely factor is a difference in the method of determining a  $\geq 80\%$  reduction in pain intensity following the FGAI. We compared the mean score from the three most provocative tests performed before the procedure to the mean pain score with these same three tests

performed again after the procedure. To our knowledge, the current study is the first to use such a specific ‘test-retest’ requirement to evaluate changes in pain intensity. Maslowski et al. required the participant to “report their baseline pain severity on a 10-cm visual analogue scale” without associating this directly to any pain provocation manoeuvres. Similarly, the patient was asked to report pain intensity “ten to fifteen minutes” after the FGAI. No mention was made as to whether or not their participants were required or allowed load to their hip to get a sense of change in pain intensity. Another factor may have been differences in the actual performance of the quadrant test. Maslowski et al. applied a compressive force through the shaft of the femur whereas this was not incorporated during our testing. Finally, participants in the Maslowski et al. study were much older (mean age 60 years; SD 13) than those in the current study (mean age 38 years; SD 12). Increasing age has been associated with an increased prevalence and severity of intra-articular pathology of the hip (Abe et al., 2000; Botser, Martin, Stout, & Domb, 2011; Kemp et al., 2014a; McCarthy et al., 2001). However, as test sensitivity increases with increasing severity of the condition of interest, this factor seems less likely to explain the higher sensitivity in the current study (Jaeschke et al., 1994a; Pewsner et al., 2004).

Whilst the sensitivity for the other ‘impingement’ tests is generally high (79% to 91%), these values are not supported by low negative likelihood ratios that would give the clinician confidence that a negative test significantly alters the post-test probability of the presence of intra articular pathology. One of these tests, the FADDIR test, has been investigated by a number of other authors and was included in a meta-analysis by Reiman et al. (2014a). Reiman and colleagues reported pooled values for sensitivity of 0.94 (95% CI 0.90 to 0.97), specificity of 0.09 (95% CI 0.02 to 0.23), positive LR of 1.02 (95% CI 0.96 to 1.08), negative LR of 0.45 (95% CI 0.19 to 1.09) and DOR of 5.71 (95% CI 0.84 to 38.86). These values were based on the findings of four original studies that investigated the accuracy of this test for diagnosing FAI and/or labral tears (using MRA as the reference standard). Reiman et al. also performed a meta-analysis of four studies that used arthroscopic findings as the reference standard and reported very similar results (sensitivity, 0.99; specificity, 0.05; positive LR, 1.04; negative LR, 0.14 and DOR, 7.82). We reported sensitivity of 0.91 (95% CI 0.77, 0.97), specificity 0.18 (95% CI 0.08, 0.34), a positive LR of 1.11 (95% CI 0.92, 1.34), a negative LR of 0.50 (95% CI 0.14, 1.84) and DOR of 2.21 for this test. In our study, a negative FADDIR reduced the probability of a PAR from 50% to 33%. However, based on the negative

LR and its wide confidence intervals, this change in probability is relatively small and is unlikely to be clinically important. Our findings reflect the conclusions of Reiman et al. in regards to this test despite the differences in reference standards.

None of the ‘end-range rotation’ tests included in the current study demonstrate diagnostic accuracy values that indicate that they have value as stand-alone tests clinically. Of these tests, the one with the highest accuracy is internal rotation in full flexion (FFIR). This test has a high sensitivity of 0.91 (0.77, 0.97) suggesting that a negative test is useful in ruling out intra-articular pathology. The probability of such pathology being present in the event of a negative FFIR test shifts from 50% to 25%. The negative LR for this test is 0.33 (0.10, 1.13) and based on the interpretation of Jaeschke et al. (1994a), this is a ‘small but sometimes important’ change in probability.

Internal rotation performed at 90 degrees of flexion (FIR) was included by Reiman et al. (2014) in their meta-analysis. Reiman et al. reported pooled values of 0.96 (95% CI 0.81, 0.99) for sensitivity, 0.25 (95% CI 0.01, 0.81) for specificity and 1.28 (95% CI 0.72, 2.27) and 0.15 (95% CI 0.01, 1.99) for positive and negative LR’s respectively. We observed a sensitivity of 0.79 (95% CI 0.63, 0.90), specificity 0.18 (95% CI 0.08, 0.34), a positive LR of 0.96 (95% CI 0.77, 1.22) and a negative LR of 1.17 (95% CI 0.44, 3.11). In our study, a positive result for the FIR *reduced* the probability of intra-articular pathology being present (from 50 to 49%) and a negative test *increased* the probability (to 54%) of this event. The estimates of accuracy from the current study indicate that the FIR test is not valuable as a stand-alone test for intra-articular pathology. Reiman et al. made similar conclusions in regards to the results of a positive test. In contrast to our findings, Reiman and colleagues reported that a negative test shifted the probability ‘notably’ (from 87% to 50%). However, these authors drew attention to the very wide confidence intervals of the negative LR and suggested that this created some doubt as to the usefulness of this test.

Of the remaining physical tests included in the present study, the test with the largest effect on post-test probability of the presence of intra-articular pathology was a positive log roll test. This finding increased the probability of a PAR to 62%. However, the positive LR of 1.6 (95% CI 0.58, 4.4) for this test indicates that the shift in probability is relatively small.

In respect to information collected from the patient history, four variables were significantly associated ( $p = <0.05$ ) with a PAR i.e. the presence of crepitus in the hip,

dominant pain in the groin, patient reported restriction of hip movement and pain whilst jogging (intensity > 2 on NPRS). Dominant pain in the groin appears to have the greatest diagnostic utility of *all* variables included in this study with a DOR of 9.18. Considering the negative LR of 0.18 (95% CI 0.06, 0.58) and sensitivity of 0.91 (95% CI 0.77, 0.97), the clinical value of this finding is where a patient does not have dominant pain in the groin. The absence of such pain drops the post-test probability of a PAR to 16%. Other researchers have demonstrated a statistically significant association between the presence of groin pain and intra-articular pathology (Altman et al., 1991; Burnett et al., 2006; Martin et al., 2008; Sutlive et al., 2008). Two of these studies examined the diagnostic accuracy of this finding (Martin et al., 2008; Sutlive et al., 2008). Sutlive and colleagues reported sensitivity of 0.29 (95% CI 0.12, 0.52), specificity of 0.92 (95% CI 0.80, 0.97), a positive LR of 3.6 (95% CI 1.2, 11.0) and negative LR of 0.78 (95% CI 0.59, 1.0) for the presence of groin pain to diagnose hip osteoarthritis. Whilst these values contrast significantly with our own, it is difficult to compare results due to major differences in methods between these studies. Of particular note, Sutlive et al. used x-rays as a reference standard to identify hip OA. However, they had no way of proving that the radiological findings were symptomatic whereas the FGAI utilised in the current study enabled us to differentiate symptomatic from asymptomatic intra-articular pathology. Martin et al. (2008) reported sensitivity of 0.59 (95% CI 0.41, 0.75), specificity of 0.14 (95% CI 0.05, 0.33), a positive LR of 0.67 (95% CI 0.48, 0.98) and negative LR of 3.0 (95% CI 0.95, 9.4) for the presence of groin pain to diagnose intra-articular hip joint pathology. Although these authors used pain response to FGAI as their reference standard, their criteria for a PAR was a  $\geq 50\%$  reduction in pain intensity, much lower than the  $\geq 80\%$  required in our study.

The presence of crepitus had a high specificity (0.91 (95% CI 0.77, 0.97)) and increased the probability of a PAR to 79% in our study. The positive LR of 3.67 (95% CI 1.12 to 11.9) suggests that this is a relatively small, but may be an important, change in probability (Jaeschke et al., 1994a). Whilst other authors (Burnett et al., 2006; Clohisy et al., 2009; Martin et al., 2008) have investigated mechanical symptoms such as catching, locking and popping, it appears that none have considered crepitus. Whilst patient reported limitation of movement and the presence of pain whilst jogging were both statistically associated with a PAR in the current study, the estimates of diagnostic accuracy reveal that these findings have poor utility as stand-alone tests.

## 5.8 Limitations

There are some limitations that should be acknowledged with the current study. Whilst the sensitivity of FGAI has been demonstrated to be very high (ranging between 91% and 97%), the level specificity is less clear (ranging between 33 and 91%) (Ashok et al., 2009; Byrd & Jones, 2004a; Dorleijn et al., 2014). There is also a chance that we have based some of our conclusions on the basis of false positive anaesthetic responses. Bogduk (2004b) has highlighted factors that might contribute to such an error, including a placebo response and bias introduced by the radiologist who administers the anaesthetic or the examiner who assesses the result. Whilst we followed Bogduk's recommendation and performed our diagnostic blocks under double-blind conditions, this does not eliminate all potential sources of error. We considered the use of placebo controlled comparative injections using both short and long lasting anaesthetic. However, as the participants in our study were patients referred for a FGAI and an MRA as a part of their diagnostic workup, this option was not appropriate.

## 5.9 Conclusion

In conclusion, we have demonstrated that there are a number of tests (including the quadrant, FADDIR and FFIR tests) that exhibit high sensitivity and might be useful in helping to screen for intra-articular pathology. Of these, the quadrant test has the highest diagnostic utility. Only the log roll test displayed characteristics that suggest that it has some utility for identifying pain that originates from within the hip joint, however, a positive result with this test alone would not be convincing evidence of a PAR. Information collected from the patient history demonstrated some of the highest diagnostic accuracy values. The presence of crepitus was highly specific and increased the post-test probability of a PAR from 50% to 79%. Patient reported loss of movement and the presence of dominant pain in the groin were sensitive and the absence of dominant groin pain decreased the probability of a PAR to just 16%.

These findings, whilst useful clinically, have raised further questions. This study has demonstrated that none of the included tests should be relied upon as a stand-alone test to either rule in, or rule out, a positive response to an intra-articular anaesthetic injection of the hip. It is clear that various combinations of tests need to be explored to see if the accuracy of the clinical examination can be improved. Also, correlation of the anaesthetic response with MRA findings may help to clarify the interpretation of test results. These questions were explored and reported in the following chapters.

## Chapter 6 Predictors of Intra-articular Pathology of the Hip

This chapter relates specifically to Question 5 of this thesis:

*What combination of findings obtained from the clinical examination of the hip best predicts a positive response to an intra-articular injection of anaesthetic into the hip joint?*

### 6.1 Introduction and background

Whilst individual test findings provide information that can help a clinician to make a diagnosis, research in the field of diagnostic reasoning suggests that experienced clinicians recognise clusters of findings that together distinguish one possible diagnosis from another. Given that there is little evidence that any one test has sufficient diagnostic utility to rule in or rule out a specific cause of hip joint pain, the clinician has no option but to try to make sense of information gained from a number of tests. Even with an extensive knowledge base regarding the various pathologies that affect the structures of the hip and the symptoms and signs that are associated with these pathologies, a differential diagnosis is often difficult (Byrd & Jones, 2001; Reiman et al., 2014b). A good understanding of the diagnostic accuracy of individual tests will help the clinician to determine the value of information obtained with each test. However, there is a substantial overlap in the symptoms and signs that relate to the various pathologies making it very difficult to be sure of the primary source of hip joint pain.

Clinical prediction rules (aka clinical decision rules) could be considered as a research based analogy to clinical expertise. These rules consist of clusters of individual test findings that have been identified by statistical analysis of predictors associated with a specific diagnosis (Beattie & Nelson, 2006; McGinn et al., 2000). They account for the individual contribution that various findings from a clinical examination make towards establishing a diagnosis. Rather than clinicians having to spend years trying to recognise patterns of findings that distinguish one diagnosis from another, this statistical analysis provides a 'short-cut' in this process.

However, newly derived clinical prediction rules (CPR's) need to be tested and validated before they are used clinically (McGinn et al., 2000). Initially, this requires

application of the rule to new cohort of patients in a clinical setting similar to that from which the rule was derived. If the rule proves to be reproducible in this cohort, further studies using a more extensive range of patients, clinicians and clinical settings should be undertaken. A rule that demonstrates a high degree of utility across different settings is unlikely to contain information that may just have been a random finding from the cohort from which it was originally derived. Ultimately, for a CPR to reach the highest level of validation, an impact analysis that demonstrates that the rule is being utilised clinically and that it has led to real health care benefits is necessary. This requires evidence of improved patient outcomes and/or reduced levels of risk associated with the patient management and/or better cost-effectiveness.

A number of clinical prediction rules have been developed to help improve decision-making in physiotherapy practice. These include those designed as diagnostic tools (Cook et al., 2010; Fritz, Piva, & Childs, 2005; Sutlive et al., 2008) as well as numerous others that help predict the likelihood of success following a specific intervention (Cleland et al., 2010; Flynn et al., 2002; Fritz et al., 2005; Hicks et al., 2005; Sutlive et al., 2008; Vicenzino et al., 2009). Unfortunately, few of these CPR's have been through the various levels of validation recommended by McGinn et al. (2000). Fritz (2009) has suggested that further research to validate existing CPR's would demonstrate a maturation of this area of research in physiotherapy. At present, only one CPR (Sutlive et al., 2008) has been developed for the purpose of improving the diagnosis of hip related pain. This rule has not been validated and the original study has a number of methodological factors that negatively influence the overall study quality (see page 177 for further detail).

Prior to the development of any new CPR, it is appropriate to consider the potential impact of any such rule given the costs, in terms of both time and resources, required to conduct a diagnostic accuracy study, to derive and validate that rule (Beattie & Nelson, 2006; Fritz, 2009; McGinn et al., 2000). Over the past decade, there has been a rapid increase in the use of medical imaging and arthroscopy in the diagnosis and management of hip pain, particularly with respect to femoroacetabular impingement (Reiman & Thorborg, 2015). However, Reiman and Thorborg (2015) suggest that there is insufficient evidence to justify the current assessment and treatment paradigm and that further high quality research is necessary to better inform current practice. Hence, the current study was undertaken in anticipation of identifying a combination of clinical

findings that might predict the presence or absence of pain arising from an intra-articular structure of the hip. A CPR with the ability to rule out symptomatic intra-articular pathology would reduce the need for expensive medical imaging and surgical exploration of painful hips. A CPR with the ability to predict the presence of such pathology would expedite appropriate imaging and management.

This chapter considers the information obtained from the tests included in the previous chapter (Chapter 5, page 118). It investigates how well various combinations of these test findings predict a positive response to the anaesthetic injection used as the reference standard for intra-articular pathology of the hip.

## **6.2 Literature review**

The following literature search was performed primarily to identify previous research that has investigated clinical prediction rules for the hip joint. A secondary aim of this search was to identify literature relevant to methodological issues that should be considered in the design and conduct of such research. The initial search was performed in July 2011, prior to the commencement of the data collection for the diagnostic accuracy study that was detailed in the previous chapter. The search strategy was saved and re-run every six months and RSS (Really Simple Syndication) feeds were set up from library databases and websites to provide alerts for additional relevant content. A final, follow-up search was performed in June 2015. Studies were included in this review if the explicit purpose was to investigate how information from individual tests and questions used in the clinical examination could be combined to enhance the overall accuracy of the examination in identifying intra-articular pathology of hip joint. Clinical prediction rules for treatment were not included. Clinical prediction rules that were based solely on laboratory or radiological findings were also excluded.

The initial search was performed using the search strategy detailed in Chapter 2. Key concepts were identified and searched separately in three main categories summarised as: 1) "hip joint" OR "hip pain" OR "groin pain" OR groin OR hip OR "femoroacetabular impingement" OR "FAI" OR labr\* OR (osteoarthritis\* N5 hip\*) OR (OA N5 hip) OR (arthritis\* N5 hip\*) OR "ligamentum teres"; 2) accura\* OR sensitivity OR specificity OR validity OR "likelihood ratio"; 3) "clinical prediction rule\*" OR "CPR" OR "clinical decision rule\*" OR "CDR" OR "predict\*" OR "gold standard" OR "reference standard".

**Identified systematic reviews**

No systematic reviews of clinical prediction rules relevant to the diagnosis of hip joint pathology were identified by this search.

**Original studies**

Three studies (Altman et al., 1991; Birrell et al., 2001; Sutlive et al., 2008) that have developed clinical prediction rules for hip joint pathology were identified by the current search. Each of these studies investigated combinations of factors that predicted the presence of osteoarthritis. A number of studies (Jung et al., 2003; Kocher, Mandiga, Zurakowski, Barnewolt, & Kasser, 2004; Kocher, Zurakowski, & Kasser, 1999) that investigated the diagnosis of septic arthritis on the basis of laboratory findings were identified but excluded.

The earliest of the relevant studies (Altman et al., 1991) was a multicentre study that examined how well information from the history and clinical examination could distinguish patients with osteoarthritis from those with other causes of hip pain (e.g. rheumatoid arthritis, trochanteric bursitis, avascular necrosis, radiculopathy). In this study, standardised assessment protocols were utilised to retrospectively assess the records of 201 people who presented with hip pain. Of these, 114 had hip osteoarthritis (mean age 64 years; SD 13) and 57 (mean age 57 years; SD 15) had ‘other causes of hip pain’. The diagnosis of osteoarthritis was initially made “by the contributing centre” but later confirmed independently by three of the authors of the study. This consensus-based diagnosis provided the reference standard for this study. All variables included in the assessment protocol were examined for a statistical association with the diagnosis of osteoarthritis. Variables with such a relationship were included in multivariate analyses to determine if a combination of variables could differentiate patients with osteoarthritis from those without. Variables included were obtained from the patient’s history, physical examination, x-rays and blood tests. These authors considered clinical prediction rules based on the ‘number of criteria present’ method and classification trees. Classification trees separate all participants into two groups based on the variable that best determines if patients have or do not have osteoarthritis. Then, these groups are further subdivided into two subgroups based on the next best variable (determined by using a ‘goodness of split’ index) (Grajski, Breiman, Di Prisco, & Freeman, 1986). This process continues until an algorithm that balances the tree size with the overall error in classification is determined (Grajski et al., 1986).

Altman and colleagues provided details of the two best CPR's they derived using the 'number of criteria present' method. Using information from the clinical examination and blood tests, they reported a sensitivity of 54% and specificity of 89% for a rule that required the presence of hip pain and at least three of the following criteria: 1) pain on internal rotation of the hip; 2) internal rotation of  $\leq 15^{\circ}$ ; 3) erythrocyte sedimentation rate (ESR) of  $\leq 20$  mm/hour; 4) morning stiffness of the hip that lasts  $\leq 60$  minutes and 5) age  $> 50$  years. A CPR rule that included radiological findings (instead of clinical findings) had a sensitivity of 89% and specificity of 91%. This rule required hip pain and at least two of the following criteria: 1) ESR of  $\leq 20$  mm/hour; 2) radiographic evidence of femoral or acetabular osteophytes and 3) radiographic evidence of joint space narrowing. Although this study formed the basis for the American College of Rheumatology criteria for the classification of patients with hip pain associated with osteoarthritis, there are a number of methodological issues that need to be taken into account when considering their findings. Perhaps most significant is the "clinical diagnosis of osteoarthritis" that was used as the reference standard in this study. These authors performed a Delphi study to develop a list of features from the patient history, physical examination and laboratory findings that would be expected to be present or not present in a patient with hip osteoarthritis. This list was used to decide the criteria for the clinical diagnosis. However, this method introduced incorporation bias into this study as the reference standard contained a number of the index tests that were then investigated in the study (including physical tests, laboratory tests and radiological findings). The reference test needs to be independent from the index tests (Fritz & Wainner, 2001). Another factor that makes it difficult to evaluate this study and to generalise its findings is that the authors did not provide details regarding how internal rotation range of movement was measured or how the cut-off point of  $15^{\circ}$  was determined.

Birrell et al. (2001) explored how various combinations of restrictions in range of movement of the hip could predict hip radiological osteoarthritis. This multicentre study included 195 patients (median age 63 years), each of whom presented at a primary care medical practice with a 'new' episode of hip pain (i.e. they were not to have previously sought treatment for their hip problem). A standardised examination protocol that included measurement of range of motion of flexion, internal and external rotation was followed. Subsequently, all participants underwent AP pelvis radiographs that were reported on by two independent observers (details of qualifications, skill or experience

were not provided). Croft's modification (Croft et al., 1990) of the Kellgren and Lawrence grading criteria was used to determine the grade of any osteoarthritic changes observed on x-ray. The relationship between radiographic abnormalities and range of movement were determined for two thresholds of osteoarthritis (mild/moderate and severe) and for each movement direction. Using cut-off points (identified by ROC analyses) of  $94^{\circ}$  for flexion and  $23^{\circ}$  for both internal and external rotation, these authors reported that a restriction of movement in any one plane was 100% sensitive for both mild/moderate and severe osteoarthritis. However, specificity for this finding was 0%. Sensitivity remained at 100% for severe osteoarthritis when restrictions in two planes were present, but dropped slightly (to 86%) for mild/moderate osteoarthritis. The highest level of specificity was associated with restrictions in three or more planes for both mild/moderate and severe osteoarthritis (93% and 88% respectively). These authors were careful not to state that their study was designed to investigate predictors of symptomatic arthritis. Instead, they explored how restricted ROM predicted *radiological evidence* of osteoarthritic changes. The findings of this study need to be considered alongside the studies (Frank et al., 2015; Jordan et al., 2009; Kim et al., 2014) that have demonstrated abnormal radiological findings in asymptomatic populations.

The most recent study that has investigated a clinical prediction rule for the diagnosis of osteoarthritis of the hip is that of Sutlive et al. (2008). This study recruited 72 patients with a primary complaint of unilateral pain in the groin, buttock or anterior thigh who were over the age of 40 (mean age 58 years; SD 11). A physiotherapy student performed a standardised history and physical examination and a second student recorded all examination findings. Tests investigated included the FABER and quadrant as well as measures of range of movement (flexion, extension, abduction, adduction, internal and external rotation). Radiographic findings of osteoarthritis were used as the reference test. These x-rays were performed, on average, 5.8 days after the index tests (SD 9.5 days). One radiologist with 15 years of experience with musculoskeletal imaging reported on all images. Both the examiners and radiologist were blinded to each other's results. These authors reported diagnostic accuracy values for the tests examined (see page 126 for detail). They examined associations between individual variables obtained through the history and physical examination with either t-tests or chi-squared tests. Potential predictors were entered into a stepwise logistic regression to determine the most accurate combination of variables that predicted the diagnosis of hip

osteoarthritis. The following five variables were identified as the best combination of predictors: 1) 'self-reported' pain with squatting; 2) lateral hip pain with active hip flexion; 3) lateral hip or groin pain with the scour test; 4) pain with active hip extension and 5)  $\leq 25^{\circ}$  of passive internal rotation. Sutlive and colleagues reported sensitivity, specificity, likelihood ratios and post-test probabilities (along with 95% CI) for various levels of compliance with this CPR i.e. a positive finding for '1 or more', '2 or more', '3 or more', '4 or more' and 'all' variables. The presence of at least '4 or more' predictors created the highest post-test probability of a correct diagnosis (i.e. 91%). However, the confidence intervals around this estimate were very wide. Hence, these authors recommended that the presence of '3 or more' predictors was a more appropriate finding to consider clinically. They reported that probability of a patient meeting this criteria having OA was 68%. The findings of this study need to be considered alongside a number of factors. Of most concern is the use of radiological findings as the reference standard for symptomatic osteoarthritis. As previously discussed, there is little evidence to support the conclusion that radiological changes consistent with osteoarthritis are necessarily symptomatic. It is clear that many people without any history of hip pain have abnormal radiological findings (Frank et al., 2015; Jordan et al., 2009; Kim et al., 2014). Whilst these authors used a reliable method of reporting radiological arthritis (Reijman et al., 2004b), they had no way of proving that the radiological findings were symptomatic. Finally, 60% of the participants in this study had concomitant low back pain. This creates some concern that pain in the groin, buttock or thigh (that constituted the main inclusion criteria) may have been somatic referred pain from the spine.

### **Summary**

There is a dearth of research that has developed clinical prediction rules based on information obtained from the clinical examination for the identification of intra-articular pain of the hip. Three studies have demonstrated that a restriction in range of movement (in particular restriction of internal rotation) is a useful predictor of such pathology. However, the degree of restriction of movement that was considered significant varied from  $15^{\circ}$  to  $25^{\circ}$ . The method of measurement was not described in one study (Altman et al., 1991) and was performed in quite different manners in the other studies (Birrell et al., 2001; Sutlive et al., 2008).

The lack of homogeneity between studies in terms of the characteristics of the included participants, the variables measured, the method of measurement and statistical analyses

employed makes comparison of the findings of these studies difficult. This, along with consideration of the various methodological issues previously discussed (such as incorporation bias and questionable choice of reference standard) makes it difficult to make any firm conclusions on the basis of these studies. It is evident that further research is necessary to better inform decision making in respect to the diagnosis of hip joint pain. Prior to reporting the results of the current study, key methodological issues that were considered in the design and conduct of this study are discussed below.

### **6.3 Methodological Considerations**

The key considerations relevant to studies that develop clinical prediction rules are those already discussed in Chapter 5. The variables used in prediction rules are typically derived from diagnostic accuracy studies. These studies provide an opportunity to investigate the relationship between a test finding and a gold or reference standard that represents the condition of interest. Provided the diagnostic accuracy study is well designed and conducted, combinations of variables investigated within that study can be evaluated to determine the optimal combination for predicting the condition of interest. However, there are additional factors that need to be considered in the development of clinical prediction rules.

The most common statistical method utilised for this purpose is binary logistic regression. This method estimates the probability of a binary response (outcome) based on predictor (or independent) variables. It does so by estimating, on a logarithmic scale, baseline odds for the outcome and an odds ratio for each predictor. These estimates are optimal in the sense that they maximise the so-called likelihood function, which guarantees unbiasedness and statistical efficiency, as the sample size grows larger.

To enhance the likelihood of developing a CPR that will be valid across a broad range of patients and clinical settings, it is crucial that all potentially important examination variables should be considered in the development of that rule (Beattie & Nelson, 2006; Laupacis et al., 1997; McGinn et al., 2000). This consideration should be made prior to data collection given that such variables need to be included in the data collection phase of the study (Nathanson & Higgins, 2008). Various authors have recommended there should be logical reasons for including each potential predictor and that they be identified through a review of relevant literature and in consultation with expert clinicians and patients (Beattie & Nelson, 2006; Cleland, Childs, Fritz, Whitman, & Eberhart, 2007; Stanton, Hancock, Maher, & Koes, 2010).

Two important consequences of developing an exhaustive list of variables relate to collinearity and the sample size of the study. Multicollinearity can be described as an overlap of information collected from two or more variables and, as such, tends to lead to overfitting when the variables are included in the model. Multicollinearity needs to be considered during the process of variable selection, and this can be done through the use of a model comparison criterion (see page 181 for more detail). Similarly, an all-inclusive list of predictor variables creates a risk of creating a CPR based on an overfitted model that may not be reliable when applied to a different patient population (Harrell, Lee, & Mark, 1996; Hosmer, Lemeshow, Sturdivant, & Ebooks Corporation, 2013). Hosmer et al. (2013) state that there has been surprisingly little work on sample size for logistic regression. However, a commonly reported ‘rule of thumb’ is that there should be a minimum of 10 ‘events’ (in the less frequent outcome category) per predictor variable (Harrell et al., 1996; Laupacis et al., 1997; Nathanson & Higgins, 2008). For example, if in a diagnostic accuracy study of 100 participants there are 20 who have the outcome of interest and 80 that do not, the maximum number of variables that should be considered in a CPR derived from this study should be two. This figure is widely regarded as a ‘guideline’ and more recent evidence suggests that the requirement can be ‘relaxed’ to between 5-9 outcome events per variable (Vittinghoff & McCulloch, 2007). In the example above, with 20 in the smallest outcome group, between two and four variables could be considered. Ideally, a large sample size will help ensure that there will be sufficient events per variable to allow consideration of multiple variables. However, if the sample size is restricted, or the prevalence of the outcome of interest is very low, researchers need to select predictor variables that are most likely to enhance the performance of the final CPR derived from the data (Nathanson & Higgins, 2008).

The most common method of reducing the number of variables for inclusion in the preliminary modelling is to only include variables with a statistically significant relationship with the outcome of interest. In this method, simple statistical tests (Fishers Exact test, chi-squared test, independent samples t-tests) are performed on all variables to determine the strength of any relationship between a variable and the outcome. Variables are selected by examining the  $p$  values from these statistical inference tests. Despite the frequent use of this method, there appears to be little consistency regarding the level of significance that is considered appropriate. Hosmer et al. (2013) recommend using a liberal  $p$  value of 0.20 or 0.25 as a screening criterion so that important variables are not left out of the regression analysis. This is based on the argument that more strict

*p* values risk excluding variables that on their own have a ‘weak’ relationship with the outcome, yet when considered alongside other variables may have a much larger effect (Nathanson & Higgins, 2008). *P* values ranging from 0.05 (Fritz et al., 2005; Tseng et al., 2006) to 0.20 (Vicenzino et al., 2010) have been used in the development of CPR’s relevant to physiotherapy. Another method of data reduction employed in some studies (Cook et al., 2010; Fritz et al., 2005; Wainner et al., 2003) is to select variables based on their likelihood ratios. These researchers considered that a test with positive likelihood ratio of  $\geq 2$  or a negative likelihood ratio of  $< 0.5$  had sufficient diagnostic utility to be included in their logistic regression analysis.

Once the list of predictor variables and model size have been determined, models produced by logistic regression analyses need to be compared. Information criteria are an appropriate means of comparing any model on a fixed data set. These criteria estimate the distance of the currently fitted model to the true model in model space, using various notions of distance. They provide a means for selecting the optimal combination of variables by measuring the relative quality of each model. In practice they are a likelihood statistics penalised for model roughness, which is largely dependent on the number of parameters in the model (the more parameters, the rougher the model).

Various information criteria including the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and a corrected version of the AIC (AICc) can be used for this purpose. Essentially these differ by employing different strengths of penalty for the number of parameters in the model. The Bayesian Information Criterion (BIC) has the reputation of underfitting (yielding overly smooth models) and the AIC is reputed to allow too much overfitting. Lying between these two is the AICc, which corrects for finite sample size (Sugiura, 1978). The AICc considers the raw fit of the model (the log-likelihood ratio statistics) and penalises it more strongly for models with more parameters. Thus, while overfitting is always possible, the probability of overfitting with a large number of parameters is essentially nil (Vandal, 2015). The lower the value of the AICc, the smaller the amount of information lost for a given model and therefore the closer the model is to the true model. Whilst this approach to model selection allows for a virtually limitless number of model forms to be assessed, the general consensus in the literature is that some pre-selection of variables is necessary and appropriate to avoid overfitting and/or the derivation of a CPR that

contains a nonsensical combination of variables (Greenland, 1989; Laupacis et al., 1997).

## **6.4 Methods and procedures of the current study**

### **6.4.1 Data collection**

The data analysed in this study was collected as a part of the diagnostic accuracy study described in the preceding chapter. Full details of the methods of data collection and analysis are described in that chapter (page 144). However, a brief overview of the methods employed follows. Consecutive patients with hip pain who consulted a medical specialist were included in the study if the specialist considered that the patient required a magnetic resonance imaging arthrogram (MRA) and fluoroscopy guided anaesthetic injection as part of their diagnostic work-up. The researcher collected baseline data (e.g. demographics, signs, symptoms, cause, past history and aggravating and easing factors) via a questionnaire and performed a number of clinical tests (e.g. pain provocation tests, range of movement tests, resisted tests). A positive test was defined as one that reproduced a 'familiar' pain of an intensity of  $\geq 2$  points on the numeric pain rating scale. Variables included in this examination were determined by considering relevant research evidence that has demonstrated diagnostic utility of specific variables. Also, in a preliminary study, the researcher consulted five medical specialists with expertise in the diagnosis and management of patients with hip joint pain (two orthopaedic surgeons, two sports physicians and a recognised specialist physiotherapist) and four patients with hip pain. In these semi-structured interviews, items from the history and clinical examination that the specialists and patients considered important were identified. Immediately after data collection the participant underwent a fluoroscopy guided anaesthetic injection and MRA. The researcher re-examined the participant immediately after the MRA, repeating all tests in the same order as prior to the procedure. The researcher was blinded to the MRA findings and the radiologist reporting the MRA was blinded to all baseline data and from the response to the anaesthetic injection. All data collection was performed on site at the private radiological practice where the guided anaesthetic injection and MRA were performed.

### **6.4.2 Data Analysis**

Preliminary analysis was performed as a part of the diagnostic accuracy study previously described (page 149). This included testing for univariate relationship with the PAR using independent samples t-test (for normally distributed continuous

variables) and Fishers Exact Test (for categorical variables). Receiver operator characteristic (ROC) curves were constructed for continuous variables with a significant relationship to the PAR, and sensitivity and specificity values were calculated for all possible cut-off points. Post-test probabilities were calculated for all variables. Additional analyses were then performed as detailed below.

### **Variable selection and data reduction**

An initial set of predictor variables to be considered for inclusion in the multiple logistic regression were first identified on the basis of their statistical association with a PAR. Initially, all variables associated with the PAR with a level of significance of  $< 0.25$  were included. This relatively liberal significance level was selected to avoid exclusion of potential predictor variables. For the same reason, despite not having a significant univariate association with the PAR, some additional variables were included in this initial set of predictors (Beattie & Nelson, 2006; Laupacis et al., 1997). Hence, variables considered important by experts in the field and those identified as having diagnostic utility by previous research (including our own diagnostic accuracy study) were included. From this initial group of predictors, a final *reduced set of predictors* was determined by excluding variables considered inappropriate to include in the logistic regression analysis (detail provided in results section 6.5.1).

### **Logistic regression analysis**

Binary logistic regression using R statistical software (R Core Team, 2015) was performed to systematically explore all possible models that included up to 7 variables from the reduced set of predictors to determine the best model for predicting a positive response to the anaesthetic (and therefore the identification of intra-articular hip joint pain). The ceiling of 7 variables as the model size was determined as we considered that larger models would not be pragmatic for clinical use. This number of variables is also consistent with the recommendation that between 5 and 10 events per variable are appropriate (Harrell et al., 1996; Laupacis et al., 1997; Vittinghoff & McCulloch, 2007). Only main effects (i.e. no interactions) were considered in the models. Each interaction between dichotomous variables acts as an extra variable and their inclusion in a model makes the clinical application of the model more complex. It was considered preferable *a priori* to retain main effect variables for clinical prediction rather than derived variables such as interactions. Their investigation as clinical predictors remains of interest.

The corrected version of Akaike Information Criterion (AICc) (Sugiura, 1978) was employed as the criterion for measuring model adequacy. The fitted models were also assessed using area under the curve (AUC). Unlike the AICc, this criterion does not penalise for finite sample size. It was employed to provide a comparison to the findings based on the AICc criterion. The best models determined by logistic regression were then considered by the author to determine their clinical applicability and utility. Based on this analysis, the 'best' model was identified for further assessment using the Statistical Package for the Social Sciences (SPSS) software, version 22 (IBM© Corporation, 2013). The overall goodness of fit of the model was assessed using the Hosmer–Lemeshow test. The usefulness of the model was also measured with the Cox and Snell  $R^2$  and the Nagelkerke  $R^2$  statistics. The overall ability of the model to distinguish between participants who had a PAR and those that did not was assessed by calculating the area under the receiver operator characteristic (AUROC) curve. Summary measures of accuracy including sensitivity, specificity and likelihood ratios were calculated.

A probability equation that allows calculation of the estimated probability of a PAR for individual participants, based on the findings from the variables/tests included in the best model identified by the logistic regression analysis, was presented. Also, a 'Screening Score' was derived from this probability equation. This screening score provides a more simplified way to interpret the findings of the test results that may be useful clinically. Finally, a rescaled version of this score was calculated (see following text for detail) and a 'fitted' screening score was computed for each participant employing the actual test results obtained from that participant.

### **Calculation of levels of positivity and a rescaled screening score**

The purpose of logistic regression is to determine the best combination of test findings that work *together* to enable calculation of the probability of a specific outcome (a PAR in this case). An accurate estimation of probability is dependent on the inclusion of all tests in the model. Hence all tests need to be performed and all test results need to be entered into the probability equation derived from the logistic regression analysis. This equation includes the regression coefficient for each variable, which expresses the contribution each variable makes to the model. This contribution is only meaningful when considered alongside the contribution of each of the other components in the model.

However, it is not uncommon practice to use the variables identified by logistic regression analysis in a different manner (Fernández-De-Las-Peñas, Cleland, Cuadrado, & Pareja, 2008; Sutlive et al., 2008; Vicenzino et al., 2010; Vicenzino et al., 2009; Wright, Cook, Flynn, Baxter, & Abbott, 2011). These authors have provided accuracy statistics for various levels of positivity based on the test results of the variables identified by logistic regression. To allow comparisons with previous relevant studies (Birrell et al., 2001; Sutlive et al., 2008), we employed this same method of analysis. Hence, we calculated accuracy statistics for various *numbers* of positive test results by constructing 2 x 2 contingency tables using the Statistical Package for the Social Sciences (SPSS) software, version 22 (IBM© Corporation, 2013). Six levels of positivity (ranging from ‘one or more’ positive tests to all six tests positive) were considered. In this analysis, the cell counts in the 2 x 2 tables at each ‘level’ include the number of participants that fit each cell at each point in time. For example, a participant that has six positive tests is included in cell counts for level 6 (six positive tests) but also cell counts for each of the other levels. Obviously if a participant has 6 positive tests, they also have 5, 4, 3, 2 or 1 positive tests. In this manner accuracy statistics can be calculated for each level. The dependent variable was the response to the anaesthetic (either positive or negative). Following accepted practice, 0.5 was added to any 2 x 2 cells that had zero counts to enable calculation of accuracy statistics (Cox, 1970). Various measures of diagnostic accuracy were calculated including sensitivity, specificity, positive likelihood ratios (LR+) and negative likelihood ratios (LR-). 95% CIs were constructed for each variable using the Confidence Interval Calculator downloaded from the Physiotherapy Evidence Database (Herbert, 2013). In this analysis, all test findings are *weighted equally* such that *any* combination of test variables is considered to have the same influence on the probability of a given outcome.

The accuracy of the levels of positivity method for predicting a PAR was compared directly to the results obtained by using the fitted screening scores. To enable this comparison, the screening score was first rescaled so that the maximum total score (i.e. when all six tests were positive) would be six, the same number of tests included in the degrees of positivity method. In this manner, the weighting that each test result has on the probability of a PAR, as previously calculated by the probability equation (see page 194) is retained.

## 6.5 Results

Sixty-eight participants were included in the study (mean age 38.2 years; SD 11.8 years and mean BMI 24.5; SD 3.0). Demographic and baseline data for this cohort were presented in Chapter 5 (page 150). There were no missing data, with results being recorded for all variables and all patients.

### 6.5.1 Reduced set of predictors

Table 6.1 provides detail of the variables initially considered for inclusion in the logistic regression analysis. This includes those that demonstrated a statistical association with a PAR with a level of significance of 0.25 or less. For continuous variables that demonstrated a statistical association with a PAR, the cut-off points identified by receiver operating characteristic (ROC) analysis (see Chapter 5, page 156 for full details) were employed to provide a categorical interpretation of these variables. Also included in this table are five variables (painful click, FADC, log roll, passive extension and resisted extension) that did not have a level of significance below 0.25. The presence of a painful click (felt deep in the groin), the 'flexion adduction compression' (FADC) and log roll tests were included because they were considered to be key indicators of intra-articular pathology by the experts consulted in preliminary interviews (see page 182 for detail). The inclusion of the log roll test was also supported by previous research that has indicated that this test is the 'most specific' test for intra-articular pathology (Byrd, 2014; Domb et al., 2009). Resisted tests are commonly used in clinical practice as pain provocation tests and were considered by the experts as an important part of a normal clinical examination of the hip. Although these tests are most commonly used to test the integrity of the musculo-tendinous unit (Cyriax, 1974), the compressive force generated by contraction of the muscle will load the underlying joint and has the potential to provoke symptomatic intra-articular pathology. For this reason, we considered it prudent to include at least one resisted test. Resisted extension was selected as the 'best' of the resisted tests given its high specificity (85%). The inclusion of resisted extension was also considered appropriate given that there was only one highly specific variable (crepitus) with a level of association with a PAR of less than 0.25. To enhance the potential of the CPR to 'rule in' intra-articular pathology another highly specific test (passive extension) was included.

A number of variables that demonstrated a statistically significant association with the PAR were excluded from the reduced set of predictors, primarily on the basis that they demonstrated poor diagnostic accuracy (defined as a positive likelihood ratio of <1

along with a negative likelihood ratio of  $> 1$ ). Three variables excluded on this basis ('external derotation test', 'sustained one leg stand' and 'tenderness on palpation of the greater trochanteric region') were also considered inappropriate for inclusion because they are tests for extra-articular pathology (primarily tendinopathy). They were included in the diagnostic accuracy study for later consideration in Chapter 7 of this thesis (prevalence of pathology identified by MRA). Pain felt when jogging was excluded, as the utility of this test is likely to be limited in that many patients with hip pain would not jog and would therefore be unable to answer this question. Similarly, only patients that played sport could report 'stopped sport' because of their hip pain. For clarity, the final variables (n=14) selected for inclusion in the reduced set of predictors are detailed in Table 6.2.

**Table 6.1 Characteristics of the initial set of variables considered for inclusion in multiple logistic regression analysis (continued on next page)**

Variable	TP	FP	FN	TN	P Value (Fishers Exact Test)	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
<sup>1</sup> Groin Pain	31	18	3	16	0.001	0.91 (0.77, 0.97)	0.47 (0.31, 0.63)	1.72 (1.23, 2.04)	0.18 (0.06, 0.58)	9.18 (2.35, 35.89)
EDT	2	11	32	23	0.011	0.06 (0.02, 0.19)	0.68 (0.51, 0.81)	0.18 (0.04, 0.76)	1.39 (1.09, 1.78)	0.13 (0.03, 0.65)
<sup>1</sup> Age $\geq$ 39	24	12	10	22	0.015	0.70 (0.54, 0.83)	0.65 (0.48, 0.79)	2.00 (1.20, 3.31)	0.45 (0.26, 0.81)	4.40 (1.59, 12.19)
<sup>2</sup> Pain jogging	22	31	12	3	0.017	0.65 (0.48, 0.78)	0.08 (0.03, 0.23)	0.70 (0.54, 0.93)	4.00 (1.23, 12.92)	0.18 (0.45, 0.70)
<sup>1</sup> SR $\downarrow$ ROM	29	23	5	11	0.023	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)
<sup>1</sup> Crepitus	11	3	23	31	0.033	0.32 (0.19, 0.49)	0.91 (0.77, 0.97)	3.67 (1.12, 11.99)	0.74 (0.58, 0.96)	4.94 (1.24, 19.76)
RAB	1	8	33	26	0.043	0.03 (0.01, 0.15)	0.76 (0.60, 0.88)	0.13 (0.02, 0.95)	1.27 (1.04, 1.54)	0.10 (0.01, 0.84)
<sup>1</sup> Quadrant	33	27	1	7	0.054	0.97 (0.85, 0.99)	0.21 (0.10, 0.37)	1.22 (1.02, 1.47)	0.14 (0.02, 1.10)	8.56 (0.99, 73.9)
RAD	7	15	27	19	0.068	0.21 (0.10, 0.37)	0.56 (0.39, 0.71)	0.47 (0.22, 1.0)	1.42 (1.01, 2.01)	0.33 (0.11, 0.96)
<sup>1</sup> IR < 41 <sup>0</sup>	27	19	7	15	0.068	0.79 (0.61, 0.90)	0.44 (0.27, 0.62)	1.42 (0.93, 2.30)	0.63 (0.38, 1.04)	3.05 (1.04, 8.89)
Stopped sport	21	28	13	6	0.104	0.62 (0.44, 0.77)	0.18 (0.07, 0.35)	0.75 (0.55, 1.02)	2.16 (1.08, 4.34)	0.34 (0.11, 1.06)
SOLSt	6	13	28	21	0.104	0.18 (0.08, 0.34)	0.62 (0.45, 0.76)	0.46 (0.20, 1.07)	1.33 (0.98, 1.81)	0.35 (0.11, 1.06)
<sup>1</sup> FFIR	31	25	3	9	0.109	0.91 (0.77, 0.97)	0.26 (0.15, 0.43)	1.24 (0.99, 1.56)	0.33 (0.10, 1.13)	3.72 (0.91, 15.22)
<sup>1</sup> BKFO < 62 <sup>0</sup>	22	15	12	19	0.144	0.65 (0.38, 0.63)	0.56 (0.38, 0.72)	1.46 (0.93, 2.30)	0.63 (0.38, 1.04)	2.32 (0.87, 6.16)
IRSt	12	19	22	15	0.144	0.35 (0.21, 0.52)	0.44 (0.29, 0.61)	0.63 (0.37, 1.09)	1.47 (0.93, 2.31)	0.43 (0.16, 1.14)
<sup>1</sup> FF	29	23	5	11	0.152	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)
RF	5	11	29	23	0.152	0.15 (0.06, 0.30)	0.68 (0.51, 0.81)	0.45 (0.18, 1.17)	1.26 (0.96, 1.65)	0.36 (0.11, 1.19)
TOP GT	6	12	28	22	0.168	0.18 (0.08, 0.34)	0.65 (0.48, 0.79)	0.50 (0.21, 1.18)	1.27 (0.95, 1.71)	0.39 (0.13, 1.21)
Giving way	5	13	27	21	0.183	0.20 (0.09, 0.38)	0.62 (0.44, 0.77)	0.54 (0.25, 1.18)	1.28 (1.05, 1.58)	0.30 (0.09, 0.97)
Painful click	12	7	22	27	0.28	0.35 (0.20, 0.54)	0.79 (0.62, 0.90)	1.71 (0.77, 3.82)	0.81 (0.63, 1.06)	2.10 (0.71, 6.25)

<sup>1</sup> Variables included in logistic regression analysis; <sup>2</sup> If pain intensity  $\geq$  2 on NPRS; TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio; EDT, External De-rotation Test; SR $\downarrow$ ROM, Self reported limitation of ROM; RAB, Resisted Abduction; RAD, Resisted Adduction; RE; IR<40<sup>0</sup>, Internal Rotation at 90<sup>0</sup> less than 40<sup>0</sup>; SOLSt, Sustained One Leg Standing; FFIR, Full Flexion Internal Rotation; BKFO<62<sup>0</sup>, Bent Knee Fall Out less than 62<sup>0</sup>; IRSt Internal Rotation in Standing; FF, Full Flexion; RF, Resisted Flexion; TOP, Tenderness on Palpation of Greater Trochanter; FADC, Flexion Adduction Compression; PExt, Passive Extension Prone; RE, Resisted Extension

Table 6.1 Continued

Variable	TP	FP	FN	TN	P Value (Fishers Exact Test)	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
<sup>1</sup> FADC	27	22	7	12	0.28	0.79 (0.63, 0.90)	0.35 (0.21, 0.52)	1.23 (0.91, 1.66)	0.58 (0.26, 1.30)	2.10 (0.71, 6.25)
<sup>1</sup> PExt	7	6	27	28	0.35	0.21 (0.10, 0.37)	0.82 (0.66, 0.92)	1.17 (0.44, 3.11)	0.96 (0.77, 1.22)	1.21 (0.36, 4.07)
<sup>1</sup> Log Roll	8	5	26	29	0.50	0.24 (0.12, 0.40)	0.85 (0.70, 0.94)	1.60 (0.58, 4.4)	0.90 (0.71, 1.13)	1.80 (0.52, 6.15)
<sup>1</sup> RE	4	5	30	29	1.00	0.12 (0.05, 0.27)	0.85 (0.70, 0.94)	0.80 (0.24, 2.73)	1.03 (0.86, 1.25)	0.77 (0.19, 3.17)

<sup>1</sup> Variables included in logistic regression analysis; TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio; EDT, External De-rotation Test; SR↓ROM, Self reported limitation of ROM; RAB, Resisted Abduction; RAD, Resisted Adduction; RE; IR<41°, Internal Rotation at 90° less than 41°; SOLSt, Sustained One Leg Standing; FFIR, Full Flexion Internal Rotation; BKFO<62°, Bent Knee Fall Out less than 62°; IRSt Internal Rotation in Standing; FF, Full Flexion; RF, Resisted Flexion; TOP, Tenderness on Palpation of Greater Trochanter; FADC, Flexion Adduction Compression; PExt, Passive Extension Prone; RE, Resisted Extension

Table 6.2 Summary of variables included in the reduced set of predictors

Variable	TP	FP	FN	TN	P Value (Fishers Exact Test)	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Groin Pain	31	18	3	16	0.001	0.91 (0.77, 0.97)	0.47 (0.31, 0.63)	1.72 (1.23, 2.04)	0.18 (0.06, 0.58)	9.18 (2.35, 35.89)
<sup>1</sup> Age ≥ 39	24	12	10	22	0.015	0.70 (0.54, 0.83)	0.65 (0.48, 0.79)	2.00 (1.20, 3.31)	0.45 (0.26, 0.81)	4.40 (1.59, 12.19)
SR↓ROM	29	23	5	11	0.023	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)
Crepitus	11	3	23	31	0.033	0.32 (0.19, 0.49)	0.91 (0.77, 0.97)	3.67 (1.12, 11.99)	0.74 (0.58, 0.96)	4.94 (1.24, 19.76)
Quadrant	33	27	1	7	0.054	0.97 (0.85, 0.99)	0.21 (0.10, 0.37)	1.22 (1.02, 1.47)	0.14 (0.02, 1.10)	8.56 (0.99, 73.9)
IR < 41°	27	19	7	15	0.068	0.79 (0.61, 0.90)	0.44 (0.27, 0.62)	1.42 (0.93, 2.30)	0.63 (0.38, 1.04)	3.05 (1.04, 8.89)
FFIR	31	25	3	9	0.109	0.91 (0.77, 0.97)	0.26 (0.15, 0.43)	1.24 (0.99, 1.56)	0.33 (0.10, 1.13)	3.72 (0.91, 15.22)
BKFO < 62°	22	15	12	19	0.144	0.65 (0.38, 0.63)	0.56 (0.38, 0.72)	1.46 (0.93, 2.30)	0.63 (0.38, 1.04)	2.32 (0.87, 6.16)
FF	29	23	5	11	0.152	0.85 (0.70, 0.94)	0.32 (0.19, 0.49)	1.26 (0.96, 1.65)	0.45 (0.18, 1.17)	2.77 (0.84, 9.12)
Painful click	12	7	22	27	0.28	0.35 (0.20, 0.54)	0.79 (0.62, 0.90)	1.71 (0.77, 3.82)	0.81 (0.63, 1.06)	2.10 (0.71, 6.25)
FADC	27	22	7	12	0.28	0.79 (0.63, 0.90)	0.35 (0.21, 0.52)	1.23 (0.91, 1.66)	0.58 (0.26, 1.30)	2.10 (0.71, 6.25)
PExt	7	6	27	28	0.35	0.21 (0.10, 0.37)	0.82 (0.66, 0.92)	1.17 (0.44, 3.11)	0.96 (0.77, 1.22)	1.21 (0.36, 4.07)
Log Roll	8	5	26	29	0.50	0.24 (0.12, 0.40)	0.85 (0.70, 0.94)	1.6 (0.58, 4.4)	0.90 (0.71, 1.13)	1.80 (0.52, 6.15)
RE	4	5	30	29	1.00	0.12 (0.05, 0.27)	0.85 (0.70, 0.94)	0.80 (0.24, 2.73)	1.03 (0.86, 1.25)	0.77 (0.19, 3.17)

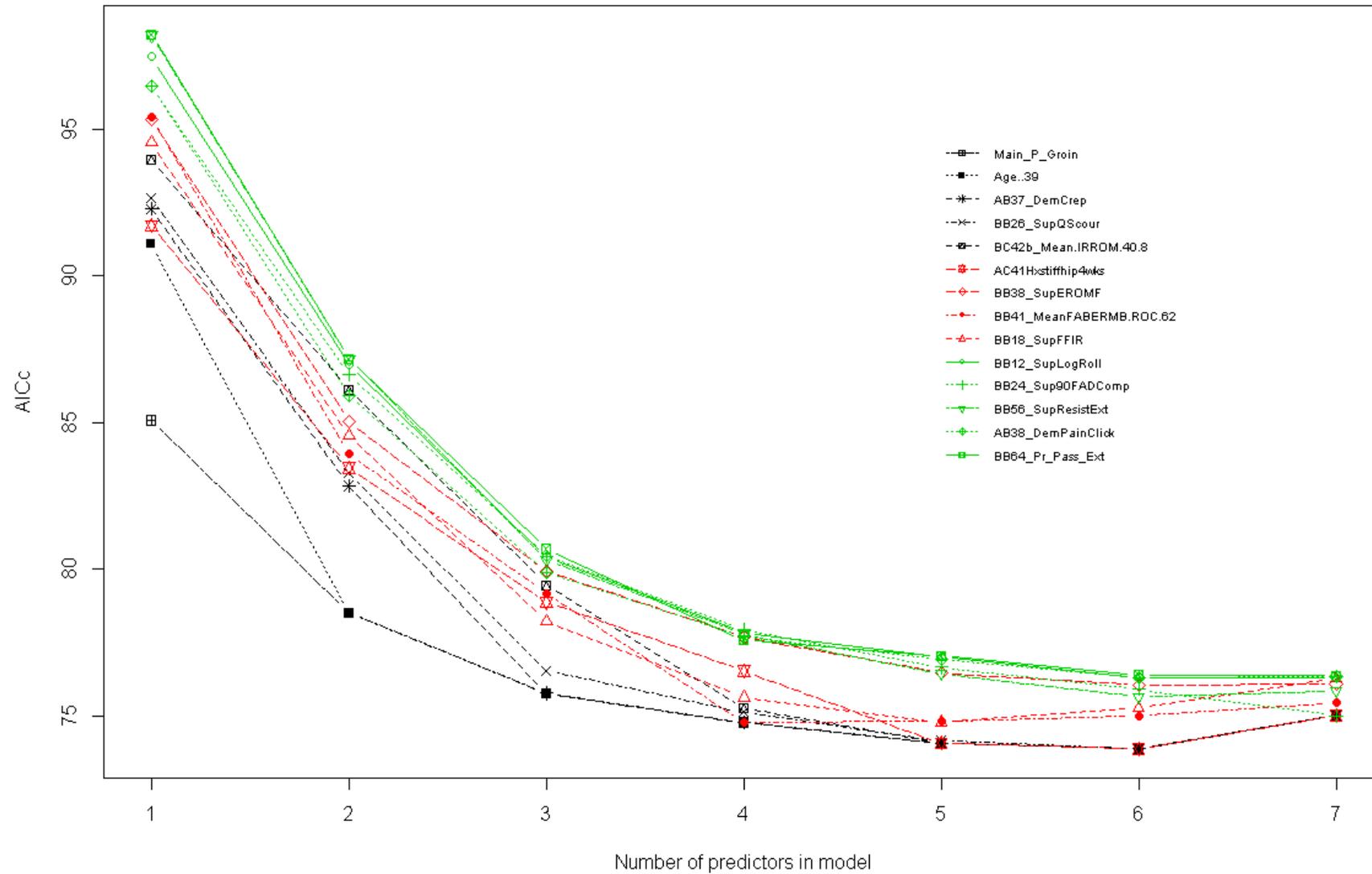
TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Diagnostic Odds Ratio.

SR↓ROM, Self reported limitation of movement; IR<41°, Internal Rotation at 90° ≤ 41°; FFIR, Full Flexion Internal Rotation; BKFO<62°, Bent Knee Fall Out < 62°; FF, Full Flexion; FADC, Flexion Adduction Compression; PExt, Passive Extension Prone; RE, Resisted Extension.

### 6.5.2 Model Selection

Figure 6.1 demonstrates the best (lowest) AICc values for each of the 14 predictor variables selected in the preliminary analysis for models of various sizes i.e. models with one to seven variables. There is an obvious trend towards a smaller (better) value of the AICc with each subsequent addition of another variable until the seventh variable is added. In this graph, the variables depicted in green are relatively poor predictors, those in black good predictors and those in red, somewhere in between. The graph clearly demonstrates that dominant pain in the groin is the best predictor regardless of the number of variables in the model. Age  $\geq 39$ , the quadrant test and the presence of crepitus also stand out as key predictors.

Table 6.3 provides detail for the best logistic regression models with, at most, seven predictors selected from the reduced set of predictors, as determined using AICc values. From this analysis ‘main pain in the groin’ was identified as the single best predictor of a PAR. This variable was retained in each of the seven best models identified by this analysis. Similarly ‘Age  $\geq 39$  years’ and ‘crepitus’ were included in all models regardless of size. The smallest AICc value for models of one predictor is 85.07. This value decreases as the model size increases up to the point where there are 6 predictors in the model (AICc=73.87). The best model with seven predictors leads to an increased AICc value (75.00) indicating that models of this size overfit the data. Hence, the best model based on AICc contains six variables. These are: 1) dominant pain groin; 2) age  $\geq 39$  years; 3) the presence of crepitus; 4) internal rotation ROM  $< 41^{\circ}$ ; 5) self-reported limited ROM and 6) positive quadrant test.



**Figure 6.1** Best value of AICc for each predictor as a function of model size

**Table 6.3 Details of AICc-optimal models for each model size**

Number of Predictors in Model	<sup>1</sup> AICc Value	Predictor variables included
1	85.07	Groin pain
2	78.5	Groin pain, Age $\geq$ 39 years
3	75.77	Groin pain, Age $\geq$ 39 years; Crepitus
4	74.77	Groin pain, Age $\geq$ 39 years; Crepitus; BKFO $<$ 62 <sup>0</sup>
5	74.06	Groin pain, Age $\geq$ 39 years; Crepitus; IR ROM $<$ 41 <sup>0</sup> ; SR $\downarrow$ ROM
6	73.87	Groin pain, Age $\geq$ 39 years; Crepitus; IRROM $<$ 41 <sup>0</sup> ; SR $\downarrow$ ROM; Positive Quadrant Test
7	75.00	Groin pain, Age $\geq$ 39 years; Crepitus; IR ROM $<$ 41 <sup>0</sup> ; SR $\downarrow$ ROM; Positive Quadrant Test; Positive FADC Test

<sup>1</sup> AICc, Corrected Akaike Information Criterion; BKFO, Bent Knee Fall Out; IR ROM, Internal Rotation Range of Movement; SR $\downarrow$ ROM, self-reported limitation of ROM; FADC, Flexion Adduction Compression

For comparison, Table 6.4 provides detail of the best models derived by using the area under the curve (AUC) as the criterion for measuring model adequacy. Here, an *increasing* AUC value indicates improved model fit, *without* compensating for potential overfitting. Despite the differences in how these criterion measure model accuracy, it is interesting to note that, regardless whether AICc or AUC is utilised, there is very little difference in the variables that are identified as useful predictors. For a model size of four variables, both AICc and AUC identified the same variables.

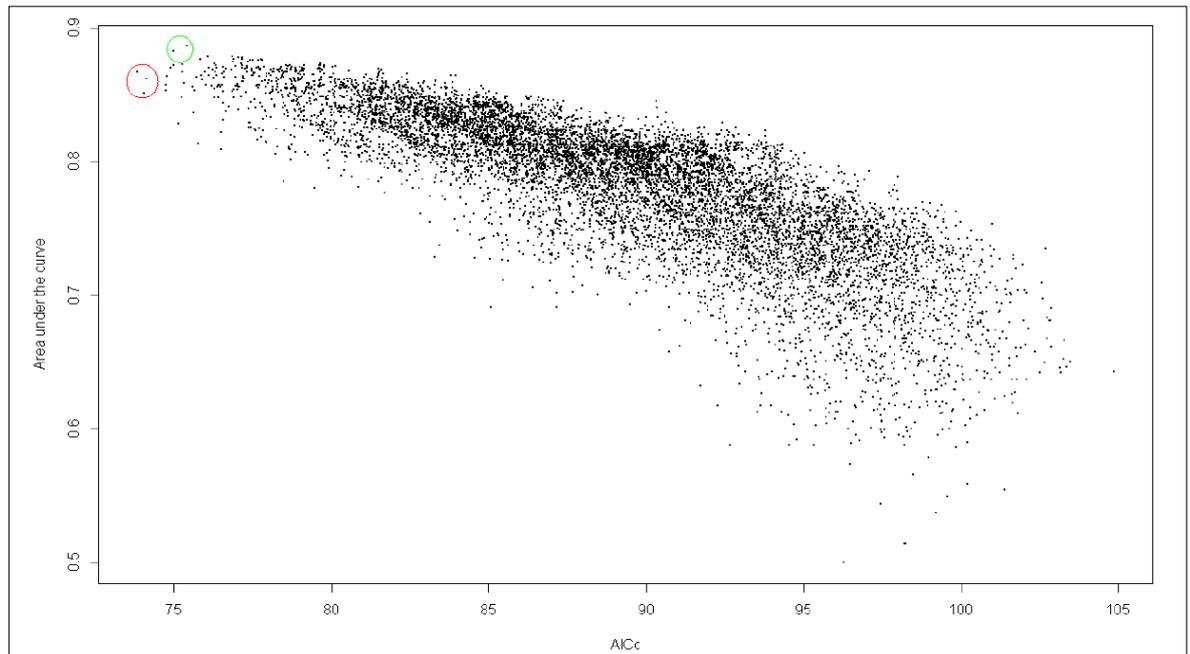
**Table 6.4 Details of AUC-optimal models for each model size**

Number of Predictors in Model	<sup>1</sup> AUC Value	Predictor variables included
1	0.691	Groin pain
2	0.785	Groin pain, Age $\geq$ 39 years
3	0.819	Groin pain, Crepitus; BKFO $<$ 62 <sup>0</sup>
4	0.857	Groin pain, Age $\geq$ 39 years; Crepitus; BKFO $<$ 62 <sup>0</sup>
5	0.870	Groin pain, Age $\geq$ 39 years; Crepitus; BKFO $<$ 62 <sup>0</sup> ; Positive FFIR
6	0.872	Groin pain, Age $\geq$ 39 years; Crepitus; BKFO $<$ 62 <sup>0</sup> ; IR ROM $<$ 41 <sup>0</sup> ; SR $\downarrow$ ROM
7	0.886	Groin pain, Age $\geq$ 39 years; Crepitus; BKFO $<$ 62 <sup>0</sup> ; IR ROM $<$ 41 <sup>0</sup> ; SR $\downarrow$ ROM Positive Quadrant Test;

<sup>1</sup> AUC, Area under the curve; BKFO, Bent Knee Fall Out; FFIR, Full Flexion Internal Rotation; IR ROM, Internal Rotation Range of Movement; SR $\downarrow$ ROM, self-reported limitation of ROM

For further comparison, Figure 6.2 plots the AUC vs. the AICc value for all 9907 models considered. In this graph, each dot represents one model. Those closest to the top left corner are the best models (highest AUC and lowest AICc). The red circle encompasses the best three models using AICc and the green circle the best two using AUC. The close association between the two model-selection criteria is demonstrated by the clustering of data points in this graph along a curve running obliquely from the top left to bottom right of the chart (best-to-worst model), with narrowing in the upper

left quadrant indicative of good consistency in the selected model. This association provides additional support for the legitimacy of the AICc as a model-selection criterion.



**Figure 6.2 Area under the curve versus AICc for each model with 7 or less predictors**

The individual contribution for each variable included in the reduced predictor set was assessed by comparing the estimated log odds ratio of the variable to the AICc value for all models in which that variable was included. This analysis demonstrated three clear tiers with dominant pain groin; age  $\geq 39$  years; the presence of crepitus; internal rotation ROM  $< 41^{\circ}$  and the quadrant test all being strong predictors. Poor predictors were the log roll, FADC, resisted extension, painful click and passive extension. The remaining variables were considered adequate predictors. Based on this analysis and consideration of the face validity of the included variables, the model with six predictor variables determined by the AICc was chosen as the optimum clinical prediction model.

### 6.5.3 Assessment and characteristics of the model

The value of the Hosmer-Lemeshow goodness of fit statistic indicated that this model fitted the data well (Chi-square 5.164,  $p = .640$ ). The values of both the Cox and Snell  $R^2$  and the Nagelkerke  $R^2$  statistics (.409 and .545 respectively) indicated that the model has value. Examination of the number of expected cases per cell in 2 x 2 tables constructed between each predictor and the outcome (PAR) established that there were sufficient numbers of cases in each category to produce valid estimates of the outcome measure. All cells had values greater than 1 and only 8% of cells had less than five.

Table 6.5 provides detail of the regression coefficient, their  $p$  values and odds ratio for each variable included in the model. The regression coefficient for all variables except SR↓ROM is close to 2. Whilst the  $p$  value of this coefficient for both SR↓ROM and the quadrant test are not significant, the predictive power of the model requires the inclusion of these variables. In fact, both the regression coefficient and the odds ratio demonstrate presence or absence of a positive quadrant test has the largest effect on the model.

**Table 6.5 Coefficients and odds ratios of the variables in model predicting a PAR**

Variable	$b$ (SE)	P value for $b$	Odds Ratio (95% CI)
Constant	-7.12 (2.16)	.001	.001
Self-reported limited ROM	1.41 (0.88)	.108	4.10 (0.73, 22.89)
Age $\geq$ 39	1.76 (0.71)	.014	5.80 (1.43, 23.46)
IR < 41 <sup>0</sup>	1.89 (0.88)	.031	6.61 (1.19, 36.86)
Groin Pain	1.89 (0.89)	.035	6.62 (1.14, 38.25)
Crepitus	2.10 (0.91)	.021	8.16 (1.36, 48.88)
Quadrant	2.17 (1.42)	.127	8.77 (0.54, 143.0)

$b$ , regression coefficient; SE, Standard Error; CI, Confidence Intervals.

These values can be entered into the following equation to calculate the estimated *probability* of a PAR for any *individual* participant:

$$Probability(p) = \frac{1}{1 + e^{-[b_0 + (b_1 X_1) + (b_2 X_2) + \dots + (b_k X_k)]}}$$

By inserting the variables included in this model, the equation becomes:

$$p = \frac{1}{1 + e^{-[b_0 + (b_1 \text{SR}\downarrow\text{ROM}) + (b_2 \text{Age}) + (b_2 \text{IR}) + (b_2 \text{GroinPain}) + (b_2 \text{Crepitus}) + (b_2 \text{Quadrant})]}}$$

Adding the associated regression coefficient converts the equation to:

$$p = \frac{1}{1 + e^{-[b_0 + (1.41 * \text{SR}\downarrow\text{ROM}) + (1.76 * \text{Age}) + (1.89 * \text{IR}) + (1.89 * \text{GroinPain}) + (2.10 * \text{Crepitus}) + (2.17 * \text{Quadrant})]}}$$

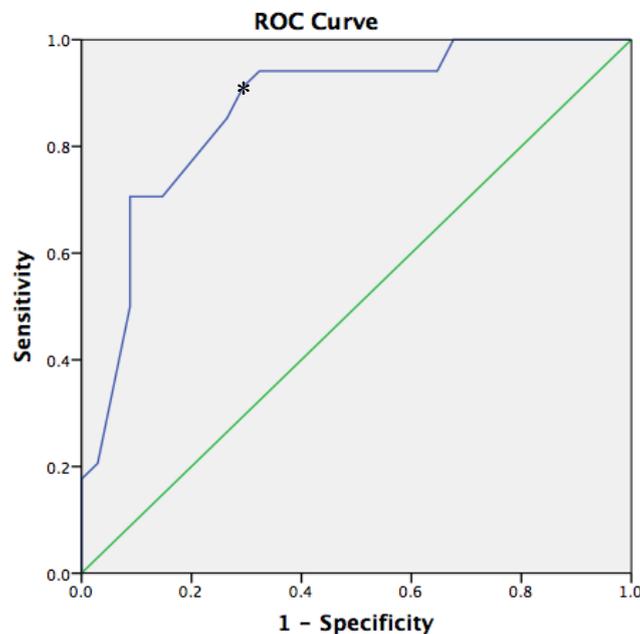
Considering the strictly monotone relationship between  $p$  and the linear predictor (under the exponential), it is appropriate to create a more simplified ‘screening score’ based on the linear predictor component of the probability equation:

$$\begin{aligned} \text{Screening Score} = & (1.41 * \text{SR}\downarrow\text{ROM}) + (1.76 * \text{Age}) + (1.89 * \text{IR}) + (1.89 * \text{GroinPain}) \\ & + (2.10 * \text{Crepitus}) + (2.17 * \text{Quadrant}) \end{aligned}$$

Finally, the score for each variable can be inserted. All variables in this model are binary, with a negative test scored as ‘0’ and a positive test as ‘1’. Hence, for a patient with all tests positive, the screening score would be calculated as follows:

$$\text{Screening Score} = (1.41 * 1) + (1.76 * 1) + (1.89 * 1) + (1.89 * 1) + (2.10 * 1) + (2.17 * 1)$$

Using this equation, a screening score was calculated for each participant in the study (i.e. an individual fitted score). Using all fitted scores, ROC analysis was performed to evaluate the discriminatory ability of the overall model. Figure 6.3 below shows the ROC curve. The AUROC was .868 (95% CI .780, .956) with a standard error of .045 and significance of <0.001. The cut-off point that maximises the sum of sensitivity and specificity (Youden-optimal cut-off point) is a screening score of 7.09. At this point, sensitivity is 91% and specificity 71%. Sensitivity was 100% for any score smaller than 4.7. Specificity at this level was only 35%. Maximum specificity (100%) occurs with a score of 9.4 or above.



**Figure 6.3** Receiver operator characteristic curve for screening scores.

\* Indicates the point of best overall accuracy for this variable.

The overall predictive power of the model was compared to the ‘null’ model where 50% of participants are assumed to be in each outcome group (PAR and NAR). The percentage of participants correctly allocated to each group by the model increases to 81% with 31 true positives, 10 false positives, 24 true negatives and 3 false negatives. Based on these cell counts, the sensitivity of the model was 0.91 (95% CIs 0.75, 0.98) and specificity 0.70 (95% CI 0.53, 0.84). Positive and negative likelihood ratios were 3.1 (95% CI 1.82, 5.27) and 0.12 (95% CI 0.04, 0.38) respectively. The positive predictive value of the model is 0.75 (95% CI 0.59, 0.87) and the negative predictive value 0.88 (95% CI 0.70, 0.97). These values indicate that the model has better utility to rule out a PAR than to rule in with high sensitivity, low negative likelihood ratios and a low number of false negatives.

#### 6.5.4 Screening score versus levels of positivity

Table 6.6 below shows the original and rescaled values of the coefficients (weightings) for each of the variables identified by the logistic regression analysis. This rescaling enables comparison of the accuracy of the levels of positivity method of analysis against the outcomes that would result from using the screening score derived from the logistic regression analysis (see page 185 for rescaling detail).

**Table 6.6 Rescaled test coefficients (weightings)**

<b>Variable</b>	<b>Original Coefficient</b>	<b>Rescaled Coefficient</b>
Self-reported limited ROM	1.41	0.75
Age $\geq$ 39	1.76	0.94
IR < 41 <sup>0</sup>	1.89	1.01
Groin Pain	1.89	1.01
Crepitus	2.10	1.12
Quadrant	2.17	1.16
	Sum of scores	11.22
		6.00 <sup>1</sup>

<sup>1</sup> Any discrepancy is due to rounding to the 2<sup>nd</sup> decimal

ROC analysis was repeated using the rescaled coefficients to determine sensitivity and specificity of cut-points based on this rescaled screening score (RSS). The cut-off point that maximises the sum of sensitivity and specificity (Youden-optimal cut-off point) is a score of 3.8. At this point, sensitivity is 91% and specificity 71%. Sensitivity was 100% for any score smaller than three. Specificity at this level was 35%. Maximum specificity (100%) occurs with a score of greater than 5. Table 6.7 provides details for the accuracy at each of 6 cut-points i.e. a RSS of 1 point or greater, 2 points or greater and so on up to 6 points. Sensitivity values range from 8% (level 6) to 98% (levels 1 & 2). Specificity ranges from 6% for level 1, to 98% for level 6 (see Table 6.7 for detail). Point estimates

for the positive likelihood ratios for levels 5 and 6 are moderate, although the confidence intervals around these estimates are relatively wide (Jaeschke et al., 1994a). The negative likelihood ratio for level 3 is moderate (0.11) and the confidence intervals around this estimate are relatively narrow. This finding is complemented by high sensitivity at this level (94%). The Youden-optimal point (i.e. a score of 3.8) demonstrates high sensitivity (91%) and a low (good) negative likelihood ratio (0.12). The largest shifts in the probability of a PAR are associated with scores of 4, 5 and 6. With these scores, this probability of a PAR increases to above 80%.

These accuracy statistics can be used to consider the utility of the levels of positivity method of decision-making. Table 6.8 provides details for the diagnostic accuracy of the six levels of positivity calculated from the test results for each of the variables identified by the logistic regression analysis (see page 185 for detail). Sensitivity values range from 9% (level 6) to 98% (levels 1 to 3). Specificity ranges from 0% for level 1 to 98% for the level 6. Point estimates of the positive likelihood ratios and specificity values are not dissimilar to those associated with the corresponding level for the RSS, except for level 4 where the levels of positivity approach under-estimates these measures. In contrast, the level of positivity method enhances the negative likelihood ratios over three levels (3-5). Similarly, this method inflates the sensitivity over these levels 4 and 5.

**Table 6.7 Accuracy statistics associated with rescaled screening score for predicting a PAR**

Level of Score	Rescaled Screening Score	TP	FP <sup>1</sup>	FN <sup>1</sup>	TN	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	<sup>2</sup> Post-test probability of a PAR in % given a positive result (CI's)
1	≥ 1.0	34	32	0	2	0.98 (0.85, 1.00)	0.06 (0, 0.21)	1.04 (0.95, 1.14)	0.24 (0, 9.73)	52 (49, 54)
2	≥ 2.0	34	26	0	8	0.98 (0.85, 1.00)	0.23 (0.11, 0.41)	1.28 (1.06, 1.56)	0.06 (0, 1.16)	57 (52, 61)
3	≥ 3.0	32	16	2	18	0.94 (0.79, 0.99)	0.52 (0.35, 0.69)	2.00 (1.38, 2.88)	0.11 (0.02, 0.45)	67 (58, 74)
4	≥ 4.0	24	5	10	29	0.70 (0.52, 0.84)	0.85 (0.68, 0.94)	4.80 (2.08, 11.09)	0.34 (0.20, 0.58)	83 (68, 92)
5	≥ 5.0	6	1	28	33	0.18 (0.07, 0.35)	0.97 (0.83, 0.99)	6.00 (0.76, 47.21)	0.84 (0.72, 0.99)	86 (43, 98)
6	6.0	3	0	31	34	0.08 (0.02, 0.24)	0.98 (0.85, 1.0)	6.09 (31, 117)	0.92 (0.83, 1.02)	86 (24, 99)
<b>Youden-optimal threshold</b>	≥ 3.8	31	10	3	24	0.91 (0.75, 0.97)	0.70 (0.52, 0.84)	3.1 ( 1.8, 5.27)	0.12 (0.04, 0.37)	76 (65, 84)

<sup>1</sup>0.5 added to cells with zero values to allow estimations of accuracy; <sup>2</sup>Probability of success based on pre-test probability of 50%; TP, True Positives; FP, False Positives; FN, False Negatives, TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Variables included in model = Groin pain, Age ≥ 39 years; Crepitus; Internal ROM <41°; SR ↓ROM, Self reported limitation of ROM; Quadrant

**Table 6.8 Accuracy statistics associated with various levels of positivity for predicting a PAR**

Level of Positivity	Number of positive clinical findings	TP	FP <sup>1</sup>	FN <sup>1</sup>	TN <sup>1</sup>	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	<sup>2</sup> Post-test probability of a PAR in % given a positive result (CI's)
1	One or more	34	34	0	0	0.98 (0.85, 1.00)	0.01 (0, 0.14)	1.01 (0.94, 1.00)	1.00 (0, 746)	50 (48, 51)
2	Two or more	34	28	0	6	0.98 (0.85, 1.00)	0.18 (0.07, 0.35)	1.20 (1.02, 1.40)	0.08 (0, 1.68)	55 (52, 59)
3	Three or more	34	23	0	11	0.98 (0.85, 1.00)	0.32 (0.18, 0.50)	1.47 (1.15, 1.84)	0.04 (0, 0.79)	60 (54, 65)
4	Four or more	32	11	2	23	0.94 (0.79, 0.99)	0.67 (0.49, 0.82)	2.90 (1.78, 4.76)	0.08 (0.02, 0.34)	74 (64, 83)
5	Five or more	17	3	17	31	0.50 (0.33, 0.67)	0.91 (0.75, 0.98)	5.67 (1.83, 17.57)	0.55 (0.39, 0.77)	85 (65, 95)
6	Six	3	0	31	34	0.08 (0.02, 0.24)	0.98 (0.85, 1.0)	6.09 (31, 117)	0.92 (0.83, 1.02)	86 (24, 99)

<sup>1</sup>0.5 added to cells with zero values to allow estimations of accuracy; <sup>2</sup>Probability of success based on pre-test probability of 50%; TP, True Positives; FP, False Positives; FN, False Negatives, TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio; DOR, Variables included in model = Groin pain, Age ≥ 39 years; Crepitus; Internal ROM <41°; SR ↓ROM, Self reported limitation of ROM; Quadrant

## 6.6 Discussion

This is the first study to have investigated the development of a CPR using information obtained from the clinical examination to help identify people who are likely to get significant pain relief as a result of an intra-articular injection of anaesthetic into the hip joint. Previous studies (Altman et al., 1991; Birrell et al., 2001; Sutlive et al., 2008) have reported that some combinations of clinical and/or laboratory findings are useful for predicting osteoarthritis of the hip. However, none of these studies used a reference standard that provides convincing evidence that the source of the pain being investigated was intra-articular in origin.

This study has provided new evidence that suggests that a combination of findings obtained from the clinical examination of a painful hip can help to predict the likelihood of a positive anaesthetic response (PAR). Six variables were included in the most parsimonious model: 1) dominant pain groin; 2) age  $\geq 39$  years; 3) the presence of crepitus; 4) internal ROM  $< 41^{\circ}$ ; 5) self-reported limited ROM (SR $\downarrow$ ROM) and 6) positive quadrant test. Whilst each of these variables has some diagnostic utility in their own right, none display characteristics that suggest that they can be used as stand-alone tests to rule in or out intra-articular pathology of the hip. Conversely, when considered together, the probability of a PAR for a patient in whom all of these variables were positive increased to 86% in this study. The findings of this study could be used by clinicians to estimate the probability of PAR for an *individual* patient by using the probability equation that we provided (see page 194). This concept has been adopted widely in other areas of medicine. For example, information collected from an individual patient can be entered into online calculators that will determine probability of that individual having a stroke or heart attack. If using the probability equation itself is not pragmatic, a more simplified method would be to use the 'screening score' that we described. Adding the regression coefficient for each positive test would give a score that appropriately 'weights' the contribution of each test to the probability of a PAR.

Although this is the appropriate way to use the results of logistic regression, it may not be pragmatic for clinicians to incorporate this additional component of decision making into their practice. If this is not considered practical, there are other ways of utilising our findings in clinical practice. The logistic regression analysis identified six variables that when considered together have predictive power for determining the likelihood of a PAR. Previous authors (Birrell et al., 2001; Sutlive et al., 2008) in this area of research

have provided accuracy statistics that describe how different numbers of positive test findings perform. Using this ‘levels of positivity’ (LOP) approach, our results demonstrate that the post-test probability of a PAR increases substantially in the presence of 4, 5 or 6 positive findings (to 74%, 85% and 86% respectively). The LOP approach is attractive from a clinical point of view in that clinicians can simply add the number of positive test findings for an individual patient and then consider the accuracy statistics we reported (in Table 6.8) for that number of positive tests. However, this approach weights the result of each test equally, which is not a true reflection of their contribution to the prediction of a PAR. As revealed by the logistic regression analysis, the contribution of each test varied, with self-reported limitation of ROM having the smallest influence and the quadrant test having the largest influence.

This study proposes a novel way to assess the accuracy of the LOP approach to decision-making. Various scores obtained by employing the rescaled screening score (RSS) (see page 185) were directly compared to the scores obtained from the use of the LOP approach. Here, the *fitted* RSS is considered the gold standard in that the actual test results for all participants were included, along with the appropriate weighting for each test (as determined by the logistic regression analysis). This comparison revealed that the performance of the LOP approach was acceptable. Whilst it underestimates some measures (e.g. the positive likelihood ratios associated with 4 or more positive tests) and enhances others (e.g. the sensitivity and negative likelihood ratios over levels 4 and 5), the overall findings are similar to those determined with the RSS.

The acceptable performance of the LOP approach as compared to the optimal RSS can be attributed to the fact that the rescaled coefficients were not very far from 1 (1 being the value attributed to each coefficient by the positivity level approach). The LOP approach can therefore be validated as a reasonable approximation of the optimal approach. However, it may be that this result was fortuitous. Rescaled coefficients much further away from 1 would tend to discredit the LOP approach, although it may be that such discrepancy is rare in optimal models. This aspect merits further investigation.

The RSS could be employed in clinical practice to enhance management decisions. Different cut-off points on this scale could be chosen depending on the ‘cost’ of making a decision based on this information. For example, a low score (say 2.0) that has high sensitivity (98%) might be chosen if it is considered important not to ‘miss’ patients who have intra-articular pathology. With this level of sensitivity, there will be very few

false negatives. Hence, a score of 2 or less helps to rule out the likelihood that the patient will experience a PAR. It is probably unnecessary, perhaps even inappropriate, to refer such patients for further investigations (e.g. an intra-articular injection, MRA or arthroscopy) if the 'risk/cost to benefit' ratio of having these procedures is considered. Alternatively, a score above 5.0 was 97% specific for a PAR in this study. For patients with such a score, referral for a guided intra-articular anaesthetic injection would provide additional information that would help to confirm if the pain is intra-articular in origin or not. This might be particularly appropriate if surgical management is being considered. The RSS provides an accurate estimate of the likelihood of the presence of intra-articular pathology and would enable more timely identification of patients who warrant further investigation and specialist treatment. Equally, referral for unnecessary medical imaging would be avoided for those that have a low probability of having such pathology.

Our results are applicable to clinical practice. The characteristics of the included patients with respect to age, nature, stage and duration of symptoms are similar to those that would seek advice from both primary and secondary health care providers like physiotherapists and sports physicians (Laupacis et al., 1997). The variables included in the CPR have face validity and are therefore likely to be considered as appropriate tests to include in the clinical examination of a painful hip (Beattie & Nelson, 2006; Laupacis et al., 1997). Both research evidence (Altman et al., 1991; Burnett et al., 2006; Clohisy et al., 2009) and the opinion of the experts consulted in this study suggests that the dominant pain is commonly felt in the groin for people with intra-articular pathology of the hip. Similarly, a loss of internal rotation (Altman et al., 1991; Birrell et al., 2001; Clohisy et al., 2009; Kemp et al., 2014b; Sutlive et al., 2008) and positive impingement signs (Burnett et al., 2006; Clohisy et al., 2009; Martin et al., 2008; Maslowski et al., 2010; Reiman et al., 2014a; Sutlive et al., 2008) have been previously associated with intra-articular pathology of the hip. The identification of age  $\geq 39$  years as a key predictor reflects the relationship between increasing age and the prevalence of intra-articular pathology that has been reported in a number of studies (Abe et al., 2000; Botser et al., 2011; Kemp et al., 2014a; McCarthy et al., 2001). Crepitus is associated with osteoarthritis and is considered an indicator of 'mechanical' hip pain (Neumann et al., 2007). Patient reported loss of movement (SR $\downarrow$ ROM) correlates with the restriction in movement associated with hip osteoarthritis (Altman et al., 1991).

The process of collecting the data necessary to consider this CPR is not burdensome. One would expect that most clinicians determine the area of pain and age of the patient as standard practice. Similarly, details regarding associated symptoms (crepitus, SR↓ROM) are typically obtained. The two physical tests (i.e. quadrant and range of internal rotation) are easily performed and are most likely already included in a hip examination by many clinicians. Only the measurement of the ROM internal rotation requires some technical expertise and measurement apparatus.

Whilst the CPR itself might be easy to employ and relevant to clinical practice, it is important that it goes through further testing and validation before it can be recommended for use. Without validation, it is difficult to determine if the variables identified in this study have wide applicability or are just a reflection of the cohort included in this study. Our intention is to perform such validation in follow up studies. Initially, the rule would be tested in a similar setting and with a similar cohort of participants. If the rule proves to be predictive in this first validation study, a much larger study that includes a broader spectrum of both examiners and participants will be undertaken. It was not possible to perform such validation studies within the timeframe of this thesis.

In respect to the statistical methods employed in the derivation of this CPR, it is notable that variables identified as the strongest predictors of a PAR by both AICc and AUC assessment criteria were very similar. This similarity increases the likelihood that these variables are important indicators of intra-articular pathology considering that they were identified by two statistics that measure the contribution of each variable to the model in quite different ways. Consequently, confidence in the prediction model is increased. It is also interesting to note that the six predictors retained in the model demonstrated a number of similar characteristics i.e. they all had a statistical association with a PAR of  $p < 0.07$ ; all bar SR↓ROM and IR  $\leq 41^0$  had positive likelihood ratios greater than 1 (with confidence intervals that did not cross 1) and all were in our 'top seven' in respect to the diagnostic odds ratio value. Similarly, the variables that were included in the logistic regression analysis that did not have a significant relationship with the PAR were not only excluded from the 'best model' but were shown to be poor predictors, regardless of the various models in which they were included. This finding provides support for the most commonly recommended and reported method of using a strong statistical association with the outcome variable as the primary means of selection of predictor variables to include in logistic regression analysis.

## **6.7 Limitations**

Two variables (crepitus and self-reported limitation in ROM) included in the CPR were self-reported variables i.e. this information was collected via the baseline questionnaire given to each participant on the day of data collection. Whilst the reliability of the physical tests and of patient reports of pain intensity were established in the preliminary studies (see Chapters 3 and 4 and White et al. (2015)), the reliability of patients reporting decreased ROM and crepitus was not determined. The relevant questions were: 1) Do you experience crepitus (grinding/creaking or similar)? and 2) During the last 4 weeks, have you noticed any limitation in the range of movement of your hip? During the completion of this questionnaire, participants were required to ask for clarification whenever they were unsure of the meaning of a question and the researcher was available to answer any such questions. We believe that these questions were accurately reported.

## **6.8 Conclusion**

This study has developed a clinical prediction rule that enables clinicians to predict the probability of a patient experiencing a positive response to an intra-articular injection of anaesthetic into the hip. Such a response is indicative of the presence of symptomatic intra-articular pathology. Whilst further research is required to validate this CPR, it may be that clinicians who consider that the rule is relevant to their patient cohort will find that it helps them to make decisions for improving management of their patients with hip pain. The RSS provides a simplified means of interpreting the combination of test findings for an individual patient in a way that appropriately weights the true contribution of each test to the probability of a PAR. The medium to long-term implications of this study may be that it leads to a significant reduction in the need for the invasive and expensive medical imaging and/or exploratory surgery that is currently associated with management of people with hip pain.

## **Chapter 7 The Prevalence and Diagnostic Utility of Abnormal Findings Reported in Patients Undergoing Magnetic Resonance Imaging Arthrogram of the Hip**

This chapter relates specifically to Question 6 of this thesis:

*What is the prevalence of abnormal findings identified by magnetic resonance imaging arthrograms in people with a painful hip and how accurately do these findings predict a positive response to an intra-articular injection of anaesthetic into the hip joint?*

### **7.1 Introduction and Background**

Since the discovery of X-rays by Wilhelm Roentgen in 1895, various forms of medical imaging have been utilised to gain additional information to aid the diagnosis and management of disease and injury. Although visual evidence of pathology is compelling, the increasing reliance placed on such investigations is worrying, considering the substantial evidence that many abnormal findings reported in patients have also been demonstrated in asymptomatic populations (Brinjikji et al., 2015; Lee et al., 2015; Register et al., 2012). Feddock (2007, p. 374) suggests that “technology seems to be replacing basic medical skills rather than complementing them.” Both expert opinion and research evidence suggest that the history and physical examination are an essential, if not the most important part of the diagnostic process (Cook, 2010; Deyo, 2013; Feddock, 2007). A consequence of pathology being identified via medical imaging is that it appears to add weight to a decision to explore or manage the hip surgically (Reiman & Thorborg, 2015). Interestingly, a poor response to a FGAI (that included a methylprednisolone acetate) has been demonstrated to be associated with a poor result from surgery (Ayeni et al., 2014b). One explanation of this finding is that the pathology identified pre-operatively and treated surgically, may not have been the source of the patient’s symptoms.

This study explores the value of magnetic resonance imaging arthrogram (MRA) in the assessment of hip joint pathology. It considers both the prevalence and diagnostic utility of abnormal findings identified with the MRA performed after the fluoroscopy-guided injection of anaesthetic that was used as the reference standard in Chapter 5. Whilst a full review of the literature relating to the prevalence and accuracy of such findings is beyond the scope of this thesis, the following literature search was performed to provide some context for the results of the current study.

### 7.1.1 Literature review

The specific aims of the review were to answer the following questions:

- What is the prevalence of ‘abnormal’ findings of intra-articular structures of the hip identified by magnetic resonance imaging *or* magnetic resonance imaging arthrogram in people *without* any history of hip symptoms?
- What is the prevalence of ‘abnormal’ findings of intra-articular structures of the hip identified by magnetic resonance imaging *or* magnetic resonance imaging arthrogram in people *with* hip pain?
- What is the prevalence of ‘abnormal’ findings of intra-articular structures of the hip identified by arthroscopy?
- What is the diagnostic accuracy of magnetic resonance imaging with respect to intra-articular pathology of the hip?

The initial search was performed using the search strategy detailed in Chapter 2. Key concepts were identified and searched separately in 4 main categories summarised as: 1) “hip joint” OR “hip pain” OR “groin pain” OR groin OR hip OR “femoroacetabular impingement” OR “FAI” OR labr\* OR (osteoarthritis\* N5 hip\*) OR (OA N5 hip) OR (arthritis\* N5 hip\*) OR “ligamentum teres”; 2) “magnetic resonance imaging” OR MRI OR “magnetic resonance arthrogram” OR MRA or “medical imaging”; 3) prevalence OR incidence OR “cross sectional studies” OR epidemiology; 4) accuracy\* OR sensitivity OR specificity OR validity OR “likelihood ratio”.

Only studies that investigated the prevalence of abnormal findings identified by magnetic resonance imaging (MRI), magnetic resonance imaging arthrogram (MRA) or at arthroscopy were included. Results are detailed below under sub-headings that reflect the above aims. Where a relevant systematic review was identified, the quality of the review was considered using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Liberati et al., 2009; Moher et al., 2009). Detail and key findings of that review are presented. This is followed by presentation of detail and findings of any relevant experimental studies published since these reviews. These studies were critiqued by the author using the QUADAS 2 checklist (Whiting et al., 2011). Appendix 24 provides a summary of this quality assessment. None of these additional studies were judged having a low risk of bias across all domains (patient selection, index test, reference test and flow and timing). However, four of the five

studies were judged as having a low level of concern regarding the applicability of the studies. Further detail of this assessment is provided in the following review, together with details of the relevant study.

### **7.1.2 Prevalence of abnormal findings of intra-articular structures of the hip identified by MR imaging in asymptomatic hips**

There is a large body of research that has demonstrated pathological/structural changes identified by medical imaging techniques of the hip (including x-ray, computed tomography, ultrasound and MRI/MRA) in people who do not have any history of hip pain or functional disability (Frank et al., 2015; Hack et al., 2010; Jung et al., 2011; Kang et al., 2010; Kumar et al., 2013; Lanyon et al., 2003; Lee et al., 2015; Register et al., 2012; Schmitz, Campbell, Fajardo, & Kadrmas, 2012; Silvis et al., 2011). A recent systematic review (Frank et al., 2015) considered 26 studies (primarily cohort or case control studies) that had investigated the prevalence of radiographic and/or MRI findings suggestive of femoroacetabular impingement (FAI). Across these studies, a total of 2114 asymptomatic hips were examined. Mean age of participants was 25.3 years (no standard deviation supplied). Frank et al. provided an appropriate title, a structured abstract and an introduction that detailed their rationale. Clear objectives and detail of eligibility criteria and sources of information were specified. Details of the search strategy, study selection, data extraction processes and analyses were included. Whilst Frank and colleagues reported the level of evidence (Sackett, 1989) of the included studies, the methodological quality of these studies does not appear to have been assessed in a formal manner. Also, individual study results were not presented. The authors discussed possible sources of bias in the original studies in the limitations section of their review and reported that there was a significant degree of both selection and detection bias across the studies.

Frank and colleagues reported the overall prevalence of CAM deformities, pincer deformities and labral injuries to be 37%, 67% and 68% respectively. These authors considered a third of all participants to be ‘athletic’ (college football or hockey players and army recruits) and reported that the prevalence of CAM deformities in this subgroup was much higher (54.8%) than that for the ‘non-athletic’ participants (23.1%). The authors concluded that the presence of abnormal findings in asymptomatic hips is relatively common, particularly in athletic populations. They recommended that “reliance on imaging alone is unwise” and that the emphasis should be placed on the

patient history and physical examination when making clinical decisions about patients with hip pain.

Another systematic review (Kwee et al., 2013) investigated the prevalence of ‘normal’ anatomical variants of the hip labrum. The authors provided an appropriate rationale and clear objectives for their review along with detail of their eligibility criteria and sources of information. Comprehensive details of the search strategy, study selection and data extraction processes were included. The methodological quality of studies included in the review was assessed using a checklist for which details of content were provided. However, this checklist does not appear to be a validated tool. An overall quality score and details for each study were provided. Results of the individual studies included in their review were clearly presented in various tables and associated text. Quality scores ranged from 14% to 71%, with the authors stating that overall quality was ‘moderate’ and that consequently the results of the review should be considered with caution. In this review, there were eight studies that included people with asymptomatic hips and 24 studies that investigated MRI findings in symptomatic hips (that were later correlated with surgical findings).

Kwee et al. reported that the methodological quality of the studies that investigated asymptomatic populations was ‘moderate’. Across these studies there were 812 participants (534 males; 1069 hips) with a mean age of 33.6 (range 10-85). The overall prevalence of labral tears was 19% and the prevalence of ‘abnormally shaped’ (rounded, flattened or teardrop) labrum was 11-16%, 13-37% and 41%, respectively. One study (Abe et al., 2000), included in the review of Kwee et al., considered the correlation between age and abnormal MRI findings of the labrum. In this study, 71 volunteers (age range 13-65; 41 females) without any history of hip pain underwent a MRI scan of their acetabular labrum to determine its shape and signal intensity. These authors reported that 96% of participants aged 20 or younger, had a normal triangular shape of their labrum. In contrast to this finding, only 62% of participants 50 years of age or older had a normal shape. This difference was statistically significant. Rounded and irregular shapes (both considered ‘abnormal’) were associated with aging, with irregular shapes occurring only in those aged > 40 years. Similarly, signal intensity increased with age, with a normal homogenous low intensity signal observed in 70% of those aged < 20 years, but in only 32% of those > 50 years of age. Abe et al. concluded that ‘abnormal’ findings of the labrum are common in asymptomatic hips. They suggested that these

were age related variations and that they should be considered when interpreting MR findings in people undergoing this procedure for hip pain.

One study (Lee et al., 2015) published since these reviews was identified. This study investigated the prevalence of labral tears and associated pathology observed on MRI in 70 young (mean age 26; range 19-41) asymptomatic volunteers. The authors reported that 46% of participants had intra-articular pathology. Labral tears were observed in 39% of participants, 14% had labral ossification, 10% had acetabular delamination, 5.7% had fibrocystic changes at the head-neck junction and one participant had a paralabral cyst. Isolated labral tears were seen in 23% of participants. Tears were associated with other intra-articular pathology in 16% of participants. Lee et al. concluded that MRI findings should not be relied upon to make a definitive diagnosis in people with hip pain, given the prevalence of abnormal findings in the asymptomatic population that they had studied.

### **Summary**

There is very clear evidence that a high percentage of people without any history of injury, pain or other symptoms associated with their hips, have ‘abnormal’ findings identified by MR imaging. The prevalence of labral tears across the studies identified in the current review ranged from 6% to 83%. The prevalence of abnormalities in the shape of the labrum ranged between 11% and 44%. Increased prevalence of abnormal findings appears to be associated with advancing age and with increasing levels of activity. Few studies have reported abnormalities of other intra-articular structures of the hip in asymptomatic participants.

### **7.1.3 Prevalence of abnormal findings of intra-articular structures identified by MRI/MRA in symptomatic hips**

No systematic reviews of studies that have investigated the prevalence of abnormal findings observed by MR imaging in people with hip pain were identified by the current review. However, a number of original studies (Kumar et al., 2013; Martin et al., 2008; Narvani et al., 2003; Neumann et al., 2007) were found. Narvani et al. (2003) reported preliminary findings from a small prospective study that included 18 consecutive patients (mean age 30.5 years; SD 8.45) who were all ‘keen sports people’ with hip pain. MRA scans were performed a mean of eight weeks after clinical assessment. One consultant radiologist, who had an ‘interest in musculoskeletal radiology’ (it is unclear if he had additional qualification in this speciality), reported on all MR images.

Technical detail of the MR equipment and procedures were supplied. Acetabular labral tears were observed in four patients (22%), eight patients (44%) had extra-articular pathology (e.g. iliopsoas oedema, tendinopathy) and six (33%) had normal scans.

Neumann et al. (2007) performed a retrospective review of clinical data of patients who had undergone hip MRA. The aim of this study was to determine the prevalence of labral tears in patients who had mechanical symptoms such as clicking, locking, giving way and sharp pain. All images were read by two 'experienced' (no detail) radiologists who used specific criteria for grading lesions of the cartilage, bone marrow edema (BME) and labral tears. Technical detail of the MR equipment and procedures were supplied. Of the 100 patients included in this study (mean age 39 years; range 17-76), 66% had a labral tear, 76% had lesions (fissuring, thinning) of the articular cartilage of the femoral head or acetabulum, and 29% had bone marrow oedema (which was always associated with cartilage defects). Osteophytes, subchondral cysts and subchondral sclerosis were seen in 32%, 23% and 22% of patients respectively. These authors concluded that mechanical symptoms are commonly associated with both labral tears and articular cartilage defects. The higher prevalence of labral pathology in this study compared to that of Narvani et al. most likely reflects the differences in severity and stage of the hip pathology of the participants between studies. Participants in the Narvani et al. study had a relatively short history of pain (1-4 months), with only six reporting mechanical symptoms.

The prevalence of labral tears observed on MRA was reported in a study (Martin et al., 2008) that examined the diagnostic accuracy of symptoms and signs collected by a clinical examination of patients with hip pain who were 'potential surgical candidates'. In this study, 49 patients (mean age 42 years; SD 15) recruited from sports physicians, were examined clinically and then underwent x-rays, a MRA and a fluoroscopy-guided injection of anaesthetic into their hip. An orthopaedic surgeon read all images. No detail was provided regarding the expertise of the surgeon or the MR equipment and procedures. Martin et al. reported that 96% of subjects had a 'definite' (Czerny stage IIA/B) or 'possible' (Czerny stage III A/B) labral tear (Czerny et al., 1999). These authors considered that the high prevalence of labral tears might have been a reflection of their criteria for inclusion i.e. potential surgical candidates.

The MRI findings of 30 people with mild to moderate hip osteoarthritis (Kellgren-Lawrence grade 2 or 3 based on weight bearing AP radiograph) were compared to 55

controls (no history of hip injury or radiological evidence of OA) by Kumar et al. (2013). Experienced board-certified musculoskeletal radiologists read all images, employing well-described criteria for grading lesions of the cartilage, labral tears and BME. Technical detail of the MR equipment and procedures were supplied. The authors did not state whether or not the radiologists were blinded to the status of the participants (OA versus control). All participants were recruited from the community setting via advertisement. These authors reported that the prevalence of the 'crossover' sign (20%) and positive posterior wall sign (39%) determined from the radiographs was the same in both groups. However, a statistically significant difference in the prevalence of acetabular cartilage lesions (30.9% of controls compared to 56.7% of OA participants) and of subchondral cysts (9.1% in the control group and 26.7% in the OA group) was identified via MRI. Although the prevalence of cartilage lesions of the femur was 54.6% in the control group and 73.3% in those with OA, this difference did not reach statistical significance. Kumar et al. also observed statistically significant associations between worsening Kellgren-Lawrence scores (i.e. more OA changes) and an increased number and severity of femoral and acetabular cartilage defects, as well as subchondral cysts and labral tears. These findings provide some evidence that the prevalence of abnormal findings is higher in people with hip pain than those without, and that prevalence increases with increasing severity of associated OA.

### **Summary**

It is difficult to compare the prevalence of abnormal findings reported across these studies because of the wide variation in the characteristics of the participants included in the respective studies and the differences in methods employed across these studies. However, this evidence suggests that labral pathology, identified by MRA, is highly prevalent both in patients with symptoms and signs severe enough to justify arthroscopy and in patients with a comparatively short history of hip pain for which arthroscopy has not been deemed necessary. The prevalence of labral tears and chondral lesions appears to be much higher in patients with osteoarthritis of the hip than those without this pathology. Similarly, patients with mechanical symptoms not only have a high prevalence of labral tears, but also commonly have associated pathology such as chondral lesions and bone marrow oedema.

#### **7.1.4 Prevalence of abnormal findings at arthroscopy**

Direct comparison of the prevalence of abnormal findings identified by MR imaging to those seen at arthroscopy is not necessarily appropriate. The population of patients that undergo surgical intervention represents a sub-group of people with hip pain. Typically, surgery is employed for patients with more severe or unresolved symptoms. Hence, the prevalence and severity of pathology in this group is likely to be different to that for people who have not undergone surgery. The following evidence is presented to provide a more comprehensive picture of the prevalence of intra-articular pathology and to enable consideration of any differences or similarity in prevalence across these different cohorts.

No systematic reviews of studies that have investigated the prevalence of abnormal findings observed during arthroscopy were identified by the current review. A landmark study in this area is that of McCarthy et al. (2001). These authors reported prevalence of labral pathology in a large cohort of patients who underwent hip arthroscopic surgery for mechanical hip symptoms ('pain localised to the groin pain which could be reproduced' by provocation tests). All pathology was identified and graded by one orthopaedic surgeon (the primary author) using specific grading criteria (detail provided). McCarthy and colleagues reported that 261 out of 436 (55.3%) patients had tears at the articular margin of the labrum. Labral tears were common even in the younger age groups, with 40% of patients younger than 30 years of age demonstrating this pathology. Labral fraying was identified in 35% of the total cohort, although the prevalence in patients over 60 years of age was 60%. Tears of the labrum were associated with serious chondral lesions in 85% of this older age group. Despite normal x-rays, 62% of all participants had cartilage lesions. Based on the Outerbridge classification system (Outerbridge, 1961), these lesions were considered serious for 32% of patients.

The prevalence of tears of the ligamentum teres identified during arthroscopy has been reported in two studies (Botser et al., 2011; Byrd & Jones, 2004b). Byrd and Jones conducted a study to investigate the clinical characteristics of people with ligamentum teres pathology. In this study, the authors reviewed a database of 271 consecutive patients (average age 28.3 years; range 15-53) who had undergone hip arthroscopy for "intractable hip pain, unresponsive to conservative measures", or with "imaging evidence of intra-articular pathology amenable to arthroscopy". Although this study focused on the association between clinical symptoms and signs and the degree of

ligamentum teres pathology, the authors reported that 41 patients were identified with lesions of the ligamentum teres. Hence, the prevalence of such pathology in this cohort was 15%. A more recent study by Botser et al. (2011) reported the prevalence of this pathology in patients who had undergone arthroscopy (all performed by the same orthopaedic surgeon). Indications for surgery were not provided. Of the 616 patients in this group, 502 (mean age 39.3 years; range 15-78) had an intraoperative examination of the ligamentum teres and were included in the analysis. Some patients had bilateral examinations. Consequently, the authors reported on 558 'hips'. Surgical technique and the system used to classify ligamentum tears were described. Ligamentum teres tears were identified surgically in 51% of cases, a much higher prevalence than that reported by Byrd and Jones. Full ruptures were observed in 3.8% of this cohort, degenerative tears in 5%, and partial tears in 42%. Also, 95% of the patients in this cohort had labral tears. Botser and colleagues reported that the average age of patients with ligamentum teres tears was significantly different than those without tears (42.3 versus 36.2 years respectively). They suggested that the high prevalence of ligamentum teres tears seen in their study was due to the inclusion of low-grade partial thickness tears, which they believed that previous authors might not have considered. Identification of ligamentum teres tears by MRA in this study was poor, with just nine being reported and with only five of these confirmed arthroscopically, demonstrating an overall sensitivity of just 1.8%.

Kemp et al. (2014a) reported the prevalence of chondropathy in 100 patients (aged between 18 and 60 years) who had undergone arthroscopy for "painful intra-articular hip pathology". Participants were a subgroup of 335 patients that had been operated on by a single surgeon. Of this larger group, 152 responded to an invitation to be included in the study. For one reason or another (e.g. 'too busy', 'moved away' and 'had surgery'), 52 of those that responded were not able to be included in the study. The study was conducted 12-24 months post-surgery as the authors wanted to correlate the degree of cartilage damage seen at surgery with patient outcomes. Prevalence of chondropathy  $\geq$  Grade 1 (Outerbridge classification) was 72%. Sixty-one percent had mild chondropathy (Grade I or II), whereas 39% had severe changes (Grade III-IV). Prevalence increased with increasing age, such that 100% of those older than 50 years had at least mild chondropathy compared to a prevalence of 61% in those younger than 35 years. These authors also reported the presence of co-existing pathology, with 20% of participants also having labral pathology, 5% having FAI and 44% having a

combination of FAI and labral lesions. Isolated chondropathy was seen in 31% of this group. These findings need to be considered along with the fact that they represent approximately one third of the original sample.

### **Summary**

As might be expected, these studies reported a high prevalence of abnormal findings in patients who have undergone surgical intervention for hip pain. The prevalence of labral pathology in these studies ranged from 35 to 95%. For ligamentum teres pathology, the prevalence ranged from 15% to 51%, although this variance may reflect differences in the classification/definition of such tears. There is a consensus of opinion that the prevalence of pathology increases with increasing age, suggesting that some abnormalities may be age-related changes.

### **7.1.5 Diagnostic accuracy of MRI/MRA**

The fact that abnormal pathology has been identified by MR imaging in both asymptomatic and symptomatic populations suggests that neither form of imaging is 100% specific. Similarly, the identification of pathology during surgery that was not observed with medical imaging performed prior to that surgery demonstrates that such imaging is not 100% sensitive. The following text summarises the evidence for the accuracy of this imaging technique. This is presented under subheadings that reflect various intra-articular pathologies of the hip.

#### **Labral Tears**

Smith et al. (2011) performed a systematic review and meta-analysis of studies that investigated the diagnostic accuracy of MRI or MRA for acetabular labral tears. The authors of this review provided an appropriate title, a structured abstract and an introduction that detailed the rationale for the review. Comprehensive detail of eligibility criteria, sources of information, search strategy and study selection were provided (including a PRISMA flow-chart). Two reviewers independently appraised included studies using the QUADAS tool (Whiting et al., 2006). These reviewers also performed data extraction separately, with any disagreement being resolved by consensus. Details of data analysis, including testing for the appropriateness of meta-analysis were reported. Study characteristics, study results and detail of the findings of the critical appraisal were provided for each of the individual studies included in the

review. There was sufficient clarity and transparency in the reporting of this systematic review to determine that it is a high quality review.

In this review, Smith and colleagues identified 19 relevant studies. Pooled sensitivity and specificity values for MRI and MRA were calculated. Also, pooled values were provided for the studies that used 1.5T MRI (i.e. those that used 1.0T or 3T were not included). With respect to all MRI studies, Smith et al. reported a pooled sensitivity of 0.66 (95% CI 0.59, 0.73) and specificity of 0.79 (95% CI 0.67, 0.91). These values were similar to those obtained by considering just the 1.5T MRI i.e. sensitivity 0.70 (95% CI 0.62, 0.77) and specificity 0.82 (95% CI 0.69, 0.94). Pooled values for MRA demonstrated that it is more sensitive, 0.87 (95% CI 0.84, 0.90), but less specific, 0.64 (95% CI 0.54, 0.74) than MRI. Smith et al. concluded that whilst MRA was more sensitive than MRI for detecting labral tears, MRI was more specific. However, they also highlighted a number of methodological limitations of the studies incorporated in their review. A common weakness was a limited description of participant characteristics and of the time between the imaging and arthroscopic procedures. Similarly, there was generally a lack of detail regarding the blinding of radiologists from the surgical findings, and the surgeons from the imaging findings. Hence, Smith and colleagues recommended that both MRA and MRI should only be considered as useful 'adjuncts' in the diagnosis of acetabular labral tears.

This cautious approach to the interpretation of abnormal findings identified by MR imaging is supported by the findings of McGuire et al. (2012). These researchers investigated the influence of experience and training of the radiologist on the reporting of such images. In this study, specialised musculoskeletal radiologists retrospectively reviewed the scans of 60 patients who had undergone hip arthroscopy. All scans had been read pre-operatively by 'general' radiologists of varying experience. The specialist radiologist was blinded to the clinical indications for the scan, the original report from the generalist, and findings from the arthroscopy. These authors reported that the overall diagnostic accuracy of the specialists was greater than that of the generalists for all four pathologies investigated and that the differences in accuracy were statistically significant for three of these pathologies. With respect to labral tears, the overall accuracy for specialists was 85% compared to 70% for the generalists. For acetabular chondrosis, overall accuracy was 79% and 28%, respectively. This trend was continued for FAI with accuracy being 82% for specialist and 59% for generalists. McGuire et al. also considered the overall accuracy of MRI and MRA separately. They reported that

accuracy was higher with MRA, regardless of the experience of the radiologist. This study demonstrates that the diagnostic utility of information gathered from MRI and MRA can vary significantly depending on the pathology being reported and the expertise and experience of the radiologist. This reinforces the importance of considering the overall clinical picture of the patient and the potential for misdiagnosis if imaging findings are considered alone.

Two relevant studies (Aprato et al., 2013; Reurink et al., 2012) have been published since the Smith et al. meta-analysis. Reurink et al. (2012) investigated the reliability and validity of MRA for detecting labral pathology. In this retrospective study, two radiologists (radiologist 'A' with 6 years experience in musculoskeletal radiology and radiologist 'B' with 3 years experience) independently assessed the MRA images. Study participants were selected on the basis of having undergone hip arthroscopy because of a 'clinical suspicion of a labral tear'. This suspicion was based on the presence of groin pain, a positive impingement test, mechanical symptoms such as locking or snapping and/or a decreased ROM. Patients were included provided they had undergone a MRA at the hospital where the study (and arthroscopy) was performed although it is unclear as to whether or not a consecutive sample was enrolled. Exclusion criteria included the presence of FAI and/or degenerative changes visible on x-ray. Ultimately, 93 patients (95 hips), from an original cohort of 141 hips, were included in the analysis. Technical and procedural details of the MRA were provided. Although standardised criteria were utilised to score the appearance of the labrum, these were not described in detail. The radiologists were blinded to the findings of arthroscopy but were aware that all participants had a clinical suspicion of a labral tear. The time interval between the MRA and arthroscopy was not provided. These authors reported that the prevalence of labral tears observed on MRA images was 96%. However, the inter-examiner reliability of the two radiologists reporting on these scans was only fair at best ( $k = 0.27$ ; 95% CI 0.02, 0.51). To determine the diagnostic accuracy of the MRA, reported findings were compared to surgical findings. The sensitivity MRA for detecting labral tears for both radiologists was 0.86 (95% CI 0.79, 0.93). However, specificity for radiologist 'A' was 0.75 (95% CI 0.33, 1.0) compared to 0.50 (95% CI 0.01, 0.99) for radiologist 'B'. Whilst the sensitivity in this study was relatively high, the negative predictive value (NPV) was very low (13%) and the confidence intervals relatively wide (4 to 31%). This reflected both the high prevalence of labral tears in this population and the relatively high number of false negatives. Consequently, Reurink et al. concluded that

the absence of labral pathology on MRA should *not* be considered as convincing evidence that there is not a labral tear. Similarly, the wide confidence intervals around the estimates of specificity suggests that MRA adds little to confirming a diagnosis of a labral tear in patients who already have a high clinical suspicion of a tear. The findings of this study should not be generalised to patients with a much lower suspicion of a labral tear e.g. those without mechanical signs.

Aprato et al. (2013) compared MRA findings to surgical findings in a very different cohort than that included in the Reurink et al. study. In this retrospective study, patients were included if they had both radiographic signs of FAI (including alpha angle  $>50^{\circ}$ , os acetabuli, cross-over sign, posterior wall sign or coxa profunda) and a clinical diagnosis of FAI (pain and decreased ROM with the impingement test). Two musculoskeletal specialist radiologists *jointly* assessed the MRA films (using standardised criteria) prior to the patients (n = 41) undergoing surgery. Technical detail of the MR equipment and procedures were supplied. The time interval between the MRA and arthroscopy was not provided. Similarly, the authors did not state if the surgeon was aware of the MR results (the index test) at the time of surgery. Aprato and colleagues reported sensitivity and specificity of MRA for labral tears of 91% and 86%, respectively. They also reported negative and positive predictive values. Whilst the NPV for labral tears was higher in this study (67%) than in the Reurink et al. study (13%), the confidence intervals were still relatively wide (30 to 90%). The positive predictive value (PPV) was 97%. Based on these results, the authors concluded that MRA was an 'efficacious' means of assessing labral injury. However, the lower bound of the confidence intervals associated with the NPV suggests that a negative MRA should not be a basis for dismissing the possibility of a labral tear. These results should be considered in light of the characteristics of the included population i.e. patients with radiological evidence of FAI and a positive impingement test.

### **Ligamentum teres pathology**

Two studies (Datir et al., 2014; Devitt et al., 2014) that have investigated the diagnostic accuracy of MR imaging for identifying ligamentum teres pathology were identified by the current search. Datir et al. compared MRI and MRA findings against surgical findings as the reference standard. In this study, the authors retrospectively reviewed the records of all patients who had undergone arthroscopic surgery for hip pain during the three-year period, immediately prior to the study. Of the 187 patients (average age 39 years; range 16-74) identified, five patients were excluded because there was no explicit

mention of the ligamentum teres being examined in the operation report. Although surgical findings were used as the reference standard, the surgeons did not follow a standardised procedure in respect to assessment of ligamentum teres pathology. This is most likely because the surgery was not performed with the intention of being a reference standard for this study. Three experienced musculoskeletal radiologists reviewed all MR images. The radiologists were blinded to the clinical history, physical examination and surgical findings and had to reach consensus in regard to the imaging findings. Standardised classification criteria were described, as were the technical details of the MR scanner and procedure. These authors provided point estimates of sensitivity and specificity, but did not include confidence intervals. They reported that MRA was superior to MRI in relation to both partial/degenerative tears and complete tears. Sensitivity and specificity for MRA were 83% and 93% for partial/degenerative tears compared to sensitivity of 41% and specificity of 75% for MRI. Sensitivity for both MRA and MRI for complete tears was 67%. Specificity for both MRA and MRI was also almost identical (100% versus 99%). Datir and colleagues concluded that both MRA and MRI are appropriate for diagnosing complete tears. However, MRA provided a better correlation with arthroscopic findings for partial or degenerative tears.

These estimates of accuracy contrast with those reported by Devitt et al. (2014), who investigated the diagnostic accuracy of MRI findings for the identification of ligamentum teres pathology. This study included 142 patients (average age 35 years; range 19-73) who had a MRI and subsequent arthroscopy for 'treatment of FAI'. A single, fellowship-trained musculoskeletal radiologist, blinded to all clinical data, prospectively reported MRI findings using standardised reporting criteria. Devitt and colleagues did not state if the surgeon was aware of the results of the MR imaging prior to surgery, or the time period between the imaging and surgical intervention. They reported that MRI was 90% sensitive and 8% specific for partial tears of this ligament. For hypertrophy of this ligament, sensitivity was 78% whilst specificity was 32%. Devitt et al. did not report confidence intervals making it difficult to determine the precision of these point estimates. There is no obvious reason why these results for MRI differ so much from those of Datir and colleagues given that the patient characteristics and study methods are so similar.

An even greater contrast was demonstrated by the previously discussed study of Botser et al. (2011) (see page 212 for detail). In this study, MRA images were obtained for 470 patients and MRI for 84. This study recruited patients from 73 different institutions. All

imaging was performed and interpreted by the radiologists working at the referring institution. Ligamentum teres tears were observed surgically in 284 cases (51%). MR imaging only identified five of these preoperatively. Four tears were observed on MRA that were not confirmed surgically. Based on these findings, Botser and colleagues calculated that the sensitivity of MR imaging for ligamentum teres pathology was 1.8%, and the specificity was 98.5%. However, although these authors provided this information, this study was *not* designed as a diagnostic accuracy study. Its primary purpose was to determine prevalence of ligamentum teres tears in patients undergoing arthroscopy. It is difficult to compare these findings to those of the previous studies (Datir et al., 2014; Devitt et al., 2014) given that Botser et al. did not provide details regarding the training and experience of the radiologists, the use of any reporting criteria, the methods of imaging or characteristics of the MR itself (e.g. field strength).

### **Chondral defects**

Two studies (Aprato et al., 2013; Sutter et al., 2014) that reported the diagnostic accuracy of MR imaging for defects of articular cartilage were identified in the current search. Aprato et al. (2013) compared MRA findings of patients with radiographic and clinical diagnosis of FAI to surgical findings (see page 216 for more detail). These authors reported sensitivity and specificity of MRA for lesions of the femoral head cartilage of 46% and 81%, respectively, and 69% and 88% for acetabular cartilage lesions.

Sutter et al. prospectively compared the diagnostic performance of MRA to MRI using surgical findings as the reference standard for articular cartilage lesions. This study included 28 patients (mean age 31.8 years, range 18-55) with a clinical suspicion of FAI, labral or cartilage defects. All participants underwent both an MRA and MRI, prior to hip arthroscopy. Two fellowship trained musculoskeletal radiologists independently rated all images so that inter-examiner reliability could be determined. The radiologists were blinded to the clinical data and surgical findings. Standardised radiological classification criteria were described, as were the technical details of the MR scanner and procedure. Cartilage defects were categorised as either 'extensive' or 'non-extensive'. Although arthroscopy was utilised as the reference standard, the authors of this study only provided very scant detail of the diagnostic/classification criteria used by the surgeon to identify and grade cartilage damage. The time period between imaging and arthroscopy was reported as a *mean* of 3.5 months. Whilst this might be considered sufficient time for the pathological status of some conditions to

change, this is not so likely for lesions of the articular cartilage. However, further progression of articular cartilage lesions during this time cannot be ruled out. Sensitivity and specificity for each category and each radiologist were reported separately. Intra-examiner agreement for reporting of acetabular cartilage defects was moderate ( $k = 0.5$ ) for MRA, but only fair for MRI ( $k = 0.40$ ). Agreement for reporting defects of femoral head cartilage was substantial for both MRA and MRI ( $k = 0.62$  and  $0.63$ , respectively). Diagnostic accuracy varied depending on categorisation of the cartilage lesions. For extensive defects of the acetabular cartilage, sensitivity was 100% for both radiologists using either MRA or MRI. Sensitivity ranged from 75% to 100% and specificity from 50% to 100% for extensive defects of the femoral head cartilage. In respect to non-extensive lesions of either the acetabulum or femoral head, sensitivity and specificity for MRA decreased for both radiologists. This suggests that less severe lesions of the cartilage are harder to identify accurately. Sutter and colleagues concluded by stating that MRA and MRI had similar levels of diagnostic accuracy for chondral damage of the femoral head but that MRA was both more reliable and more accurate for acetabular cartilage defects.

### **Summary of diagnostic accuracy evidence**

There appears to be a wide range of estimates of the diagnostic accuracy of MR imaging for the pathologies investigated across these studies. This variation most likely reflects differences in the characteristics of the patients included in the studies, in terms of the stage and severity of their pathology, as well as associated symptoms (e.g. the presence or absence of mechanical symptoms). Similarly, a lack of homogeneity in imaging methods, differences in technical specifications of scanners and differing radiological criteria for categorising or grading pathology, will most likely have contributed to the contrasting results. Overall, there is a general trend that suggests that MRA is superior to MRI for the pathologies included in this review. However, it is also evident that there are some issues with the inter-examiner reliability of radiologists when reporting MR images and that the expertise of the radiologist is an important influencing factor.

#### **7.1.6 Review summary**

Despite the lack of homogeneity of these studies, it is clear that 'abnormal' findings are relatively common in hip joints, regardless of the presence or absence of symptoms, the severity or duration of symptoms, patient age, the method of identification or any other variable. For this reason, there is a general consensus that imaging findings alone

should not be relied upon to make a diagnosis or to determine management (Deyo, 2013). Medical imaging should be considered an additional, not a definitive piece of information and only utilised alongside information from the patient's history and clinical findings.

It appears that no study has investigated the prevalence of abnormal findings observed on MR images in patients with hip pain, who have subsequently undergone an intra-articular injection of anaesthetic into the painful hip. Such a study would allow a direct comparison of the value of abnormal imaging findings in people who have a significant reduction in symptoms following this procedure, to those that do not experience any change. Whilst there are a number of studies that have used surgical findings as the reference standard to determine the diagnostic accuracy of medical imaging, the evidence that abnormal findings are highly prevalent in asymptomatic populations creates some doubt that surgical findings are the most appropriate means of determining if intra-articular pathology is, or is not, causing pain.

Hence, the aims of the following study were to: (1) determine the prevalence of abnormal findings identified by MRA in a cohort of patients with hip pain who have not been subject to hip surgery, and (2) to determine the diagnostic accuracy of abnormal findings identified by MRA using the response to a fluoroscopy guided, intra-articular injection of anaesthetic as the reference standard.

## **7.2 Methods and procedures of the current study**

Data analysed in this chapter was collected at the same time and from the same cohort of patients included in the diagnostic accuracy study (Chapter 5). Methods for that study were detailed in that chapter and do not warrant replication here (see page 144 for detail). However, a brief overview follows. Consecutive patients with hip pain who consulted one of the medical specialist associated with this study were referred for inclusion, provided that the specialist considered the patient required a magnetic resonance imaging arthrogram and fluoroscopy guided anaesthetic injection as part of their diagnostic work-up. The researcher collected baseline data (e.g. demographics, signs, symptoms, cause, past history and aggravating and easing factors) via a number of questionnaires and performed a number of clinical tests (e.g. pain provocation tests, range of movement tests, resisted tests). Immediately after data collection, the participant underwent a fluoroscopy guided anaesthetic injection and MRA. The researcher re-examined the participant immediately after the MRA, repeating all tests in

the same order as prior to the procedure. Sixty-eight participants were included in the study (mean age 38.2 years; SD 11.8 years and mean BMI 24.5; SD 3.0). The data obtained specifically for the current chapter and how it was analysed is presented below.

### **7.2.1 Data Collection**

Participants underwent MRA within 30 minutes of the fluoroscopy guided anaesthetic injection. Images were performed on a 3.0 Tesla scanner (Philips Achieva, Eindhoven, The Netherlands) using dual INVIVO (Gainesville, Florida) SENSE large and medium flex surface coils wrapped around the affected hip and secured with a soft Velcro strap. Images were obtained with the patient lying supine, the hip positioned in internal rotation and secured with the aid of sandbags on the knee and ankle to generate a standard version of the femoral neck across the study cohort.

The imaging protocol consisted of the following sequences; all acquired as small field of view images (130-150mm) localised to the affected hip. Slice thickness 3.0-3.5mm, interslice gap 10%.

- SPAIR (SPectral Attenuated Inversion Recovery) fat suppression
- SPIR (Spectral Presaturation with Inversion Recovery) fat suppression
- T2W Coronal and Axial SPAIR (TR 3800ms TE 60ms)
- T1W Axial, Coronal, Sagittal SPIR (TR 750ms TE 10ms)
- PDW Coronal (TR 2400ms TE 30ms)
- T1W Coronal (TR 700ms TE 10ms)
- T1 Axial Oblique (TR 700ms TE 10ms)
- T1W Radial SPIR (TR 820ms TE 9ms)

A specialised musculoskeletal radiologist (with 30 years of experience), blinded to the results of the clinical examination and effect of the fluoroscopy guided anaesthetic injection, reported on all MRA images using standard imaging criteria and a standardised reporting form developed specifically for the purpose of this study (see Appendix 25).

### 7.2.2 Data Analysis

Frequency and percentages of imaged pathology were calculated. All variables were examined for their univariate relationship with a positive response to the anaesthetic injection (see below), using Fishers exact test. Two by two contingency tables were constructed using the Statistical Package for the Social Sciences (SPSS) software, version 22 (IBM© Corporation, 2013) to examine the diagnostic accuracy of individual MRA findings. The dependent variable was the response to the anaesthetic (either positive or negative). A mean pain intensity score was calculated from the individual scores of the three most provocative tests performed prior to the fluoroscopy guided anaesthetic injection. These same tests were repeated and rescored after the MRA. A positive anaesthetic response (PAR) was defined as an 80% (or greater) reduction in this mean score post-anaesthetic. Various measures of diagnostic accuracy were calculated including sensitivity, specificity, positive likelihood ratios (LR+) and negative likelihood ratios (LR-), positive and negative predictive values and diagnostic odds ratios (DOR). Ninety-five percent confidence intervals were constructed for each variable using the Confidence Interval Calculator downloaded from the Physiotherapy Evidence Database (Herbert, 2013).

### 7.3 Results

Demographic and baseline data for this cohort were presented in Chapter Five (see page 150 for detail). Of the 68 participants who received the fluoroscopy guided anaesthetic injection, all but one had a MRA. This person (a female aged 47) withdrew from the study because of 'last minute' concern regarding claustrophobia associated with being in the MR scanner. She had a positive anaesthetic response (PAR), reporting a 100% reduction in pain intensity.

The prevalence of abnormal findings identified on the MRA images is reported in Table 7.1. This table provides detail for all participants as a group as well as separate data for participants who had either a positive or a negative response to the anaesthetic. A very high prevalence of abnormalities of the labrum was observed with only two participants having a 'normal' appearance of this structure. Tears were observed in 95% of participants and fraying in 41%. Similarly, there was relatively high prevalence of CAM lesions (69%), cartilage fissuring (63%) and cartilage thinning (54%). Femoral and acetabular cysts, and subchondral bone oedema were observed in approximately a quarter of all participants. The most commonly reported extra-articular pathologies were distension of the trochanteric bursa (37%) and tendinopathy of the gluteus minimus

(36%) and medius tendons (25%). Tears of these tendons were present in nearly a fifth of all participants. Despite the high prevalence of intra-articular pathology observed across all participants, the prevalence in participants that had a positive response to the anaesthetic was essentially the same as those that had a negative response. The only abnormality that demonstrated a statistically significant difference between PAR and NAR groups was the presence of subchondral bone oedema.

Table 7.2 provides detail about the diagnostic accuracy of individual findings from the MRA. Sensitivity ranged from 0.44 (95% CI 0.23, 0.67) for the presence of a femoral cyst to 1.0 (95% CI 0.34, 1.0) for the presence of a normal labrum. Specificity was close to 50% for almost all findings with a range from 0.33 (95% CI 0.06, 0.79) for the presence of a labral tear to 0.60 (95% CI 0.40, 0.77) for the presence of cartilage pathology. The presence of subchondral bone oedema demonstrated the highest overall diagnostic accuracy with a DOR of 3.75. Three findings (acetabular cysts, subchondral bone oedema and rupture of the ligamentum teres) had positive LR's greater than 1 and 95% confidence intervals that did not cross 1.

**Table 7.1 Prevalence of reported pathology PAR versus NAR (n=67)**

	All Cases % <sup>1</sup> (n)	PAR Group % <sup>2</sup> (n)	NAR Group % <sup>2</sup> (n)	p-value Fishers Test
<b>Bony Pathology</b>				
CAM lesion	69 (46)	70 (23)	68 (23)	1.0
Alpha Angle > 55 <sup>0</sup>	30 (20)	27 (9)	32 (11)	0.8
Osteophyte	34 (23)	33 (11)	35 (12)	1.0
Acetabular Cyst	24 (16)	33 (11)	15 (5)	0.1
Femoral Cyst	24 (16)	21 (7)	26 (9)	0.8
Subchondral Bone Oedema	22 (15)	33 (11)	12 (4)	0.043*
Loose Body	0 (0)	0 (0)	0 (0)	
<b>Cartilage Pathology</b>				
Normal	37 (25)	30 (10)	44 (15)	0.3
Thinning	54 (36)	61 (20)	47 (16)	0.3
Fissuring	63 (42)	70 (23)	56 (19)	0.3
<b>Labral Pathology</b>				
Normal	3 (2)	6 (2)	0 (0)	0.2
Fraying	61 (41)	67 (22)	56 (19)	0.4
Tear	95 (64)	94 (31)	97 (33)	0.6
<b>Bursa</b>				
Distended Trochanteric Bursa	37 (25)	30 (10)	44 (15)	0.3
Distended IP Bursa	15 (10)	18 (6)	12 (4)	0.5
<b>Tendon Pathology</b>				
Gluteus Minimus Tendinopathy	36 (24)	36 (12)	35 (12)	1.0
Gluteus Minimus Tear	18 (12)	18 (6)	18 (6)	1.0
Gluteus Medius Tendinopathy	25 (17)	27 (9)	24 (8)	0.8
Gluteus Medius Tear	18 (12)	18 (6)	18 (6)	1.0
Rectus Femoris Tendinopathy	1.5 (1)	0 (0)	3 (1)	1.0
Rectus Femoris Tear	1.5 (1)	0 (0)	3 (1)	1.0
Iliopsoas Tendinopathy	3 (2)	6 (2)	0 (0)	0.2
Iliopsoas Tear	0 (0)	0 (0)	0 (0)	1.0
<b>Ligamentum Teres Pathology</b>				
Normal	72 (48)	64 (21)	80 (27)	0.2
Swollen	19 (13)	21 (7)	18 (6)	0.8
Split	19 (13)	24 (8)	15 (5)	0.4
Rupture	1.5 (1)	3 (1)	0 (0)	0.5

PAR, Positive anaesthetic response; NAR, Negative anaesthetic response; n, number of individuals;

<sup>1</sup> percent of total sample; <sup>2</sup> percent of subgroup; \* Significant difference between PAR and NAR groups (p<0.05)

**Table 7.2 Diagnostic accuracy of imaging findings (n = 67)**

Structure	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)	DOR (95% CI)
Bony Pathology									
Alpha angle > 55 <sup>0</sup>	9	24	11	23	0.45 (0.25, 0.65)	0.49 (0.35, 0.62)	0.88 (0.50, 1.54)	1.12 (0.69, 1.84)	0.78 (0.27, 2.24)
CAM lesion	23	10	23	11	0.50 (0.35, 0.64)	0.52 (0.30, 0.74)	1.05 (0.62, 1.79)	0.95 (0.67, 1.39)	1.10 (0.39, 3.09)
Osteophyte	11	22	12	22	0.48 (0.29, 0.67)	0.50 (0.36, 0.64)	0.96 (0.57, 1.60)	1.04 (0.64, 1.7)	0.92 (0.33, 2.51)
Acetabular cyst	11	22	5	29	0.69 (0.44, 0.85)	0.57 (0.43, 0.69)	1.59 (1.01, 2.52)	0.55 (0.25, 1.18)	2.90 (0.88, 9.57)
Femoral cyst	7	26	9	25	0.44 (0.23, 0.67)	0.49 (0.36, 0.62)	0.86 (0.46, 1.60)	1.15 (0.69, 1.92)	0.75 (0.24, 2.31)
Subchondral bone oedema	11	22	4	30	0.73 (0.48, 0.89)	0.57 (0.44, 0.70)	1.73 (1.12, 2.69)	0.46 (0.19, 1.10)	3.75 (1.05, 13.35)
Cartilage Pathology									
Thinning	20	13	16	18	0.57 (0.39, 0.70)	0.58 (0.40, 0.73)	1.33 (0.80, 2.20)	0.76 (0.48, 1.23)	1.73 (0.65, 4.46)
Fissuring	23	10	19	15	0.55 (0.40, 0.69)	0.60 (0.40, 0.77)	1.37 (0.79, 2.38)	0.75 (0.47, 1.20)	1.82 (0.66, 4.96)
Labral Pathology									
Normal	2	31	0	34	1.00 (0.20, 1.0)	0.52 (0.40, 0.64)	2.10 (1.63, 2.70)	0.0	-
Fraying	22	11	19	15	0.54 (0.39, 0.68)	0.58 (0.39, 0.74)	1.27 (0.74, 2.16)	0.80 (0.50, 1.28)	1.58 (0.58, 4.25)
Tear	31	2	33	1	0.48 (0.36, 0.60)	0.33 (0.06, 0.79)	0.73 (0.31, 1.68)	1.55 (0.30, 7.8)	0.47 (0.04, 5.44)
Ligamentum Teres Pathology									
Swollen	7	26	6	28	0.54 (0.29, 0.76)	0.52 (0.39, 0.64)	1.12 (0.63, 1.98)	0.89 (0.47, 1.69)	1.26 (0.37, 4.23)
Split	8	25	5	29	0.62 (0.35, 0.82)	0.54 (0.40, 0.66)	1.33 (0.79, 2.23)	0.72 (0.34, 1.48)	1.86 (0.54, 6.40)
Rupture	1	32	0	34	1.00 (0.20, 1.0)	0.51 (0.39, 0.63)	2.06 (1.61, 2.64)	0.0	-

TP, True Positives; FP, False Positives; FN, False Negatives; TN, True Negatives; CI, Confidence Intervals; LR+, Positive Likelihood Ratio; LR-, Negative Likelihood Ratio;

DOR, Diagnostic Odds Ratio; -, Incalculable due to zero cell values.

## 7.4 Discussion

The current study has demonstrated a high prevalence of abnormalities in this group of people with symptomatic hips. Our results are largely consistent with previous research (Kumar et al., 2013; Narvani et al., 2003; Neumann et al., 2007) that has investigated the prevalence of abnormal findings observed via MR imaging of people with hip pain. The prevalence of cartilage pathology (63%), bone marrow oedema (22%), osteophytes (34%) and subchondral cysts (24%) that we observed is almost identical to that reported by Neumann et al. (2007) i.e. 76%, 29%, 32% and 23%. Similarly, Kumar et al. identified cartilage pathology in 65% and subchondral cysts in 27% of their participants. The consistency in prevalence across these studies is interesting given the different inclusion criteria for each study. Neumann and colleagues only included patients with mechanical symptoms (such as clicking, locking and sharp pain) whereas Kumar et al. only included patients with radiographic evidence of osteoarthritis. No such restrictions were employed in the current study. However, many of our participants had both mechanical symptoms and evidence of osteoarthritis. Twenty-eight percent of our participants reported painful clicking and 44% experienced painless clicking. Thirty percent reporting giving way or locking and over one-third of our participants had osteophytes. Given these findings, it may be that the nature and stage of the pathologies included in these previous studies was not substantially different to our own.

Labral pathology was identified in 95% of our participants, an almost identical figure to that reported by Martin et al. (2008). This prevalence is slightly higher than the 88% observed in the patients included in the study of Kumar et al. and moderately higher than the 66% reported by Neumann et al. In contrast, one study (Narvani et al., 2003) reported a very low prevalence with just 22% of participants having labral pathology. This low prevalence may be a reflection of the studies narrow inclusion criteria ('keen sports people' with hip pain) and the small number of participants ( $n = 18$ ).

In our study, the presence of bone marrow oedema was the only pathological finding that demonstrated a statistically significant difference in prevalence between the participants who had a PAR and those that did not. This is interesting given that previous research (Kumar et al., 2013) has demonstrated that the presence of

acetabular cartilage defects, bone marrow oedema and subchondral cysts is associated with worse self-reported disability and that acetabular cartilage defects are associated with increased pain. Similarly, Neumann et al. demonstrated that the grade of cartilage loss was correlated with the grade of both labral tears and bone marrow oedema. Hence, it might have been expected that the prevalence of these intra-articular pathologies would have been higher in the patients who reported pain relief after the intra-articular anaesthetic injection. Whilst this was the case for bone marrow oedema, it was not so for the other pathologies. Our findings suggest that the presence of these abnormalities on MRA should not be considered as evidence that the source of pain has been identified.

This study has also demonstrated that the diagnostic utility of individual abnormal findings identified by MRA is questionable. The presence of subchondral bone oedema demonstrated the highest overall diagnostic accuracy with a DOR of 3.75. However, the sensitivity, specificity and likelihood ratios associated with this finding suggest that it is unlikely to be useful clinically. The high sensitivity (100%) and low negative LR for a ruptured ligamentum teres indicate that the absence of this finding has utility for ruling out the ligamentum teres as a cause of an individual patient's hip pain. Point estimates for positive likelihood ratios for three conditions (acetabular cyst, subchondral bone oedema and ruptured ligamentum teres) were larger than one and the associated 95% confidence intervals did not contain 1. Whilst this suggests that the presence of these conditions increases the likelihood of a PAR, the values are small. Overall, our findings demonstrate that no individual structural abnormality, identified by MRA, has sufficient diagnostic accuracy to warrant ruling in that structure as the cause of a patient's pain.

Our findings contrast with some of the previous research that has investigated the diagnostic accuracy of MRA for labral pathology, using surgical findings as the reference standard. Smith et al. (2011) performed a meta-analysis of such studies and reported a pooled sensitivity of 87% and specificity of 64%, compared to the sensitivity of 48% and specificity of 33% that we observed. Two studies (Aprato et al., 2013; Reurink et al., 2012), published since the Smith et al. systematic review, reported similar estimates of accuracy to each other (sensitivity ranging from 86% to 91% and specificity from 50% to 86%). Both Smith et al. and Reurink et al.

recommended that MRA findings alone should not be considered as sufficient evidence to either rule in or rule out intra-articular pathology. In making these recommendations, Smith et al. took into consideration the limitations of many of the studies in their review including small sample sizes, lack of blinding and insufficient description of participant characteristics. Similarly, Reurink and colleagues considered that relatively high number of false negatives, the high prevalence of labral tears and the wide confidence intervals around the point estimates of accuracy in their study limited the utility of MRA in a population that already had a high suspicion of a labral tear.

Our results support those of Martin et al. (2008) who performed a very similar study to our own. In this study, patients with hip pain who were ‘potential surgical candidates’ underwent a MRA and fluoroscopy guided anaesthetic injection. These authors defined a positive response to anaesthetic injection as a > 50% reduction in pain and reported that the presence of labral tears was *not* correlated with a positive response. They concluded that labral tears observed on MRA might not be contributing to the patient’s pain and that other structures and pathologies should be considered. It is interesting that even at this much lower cut-off point for a positive anaesthetic response, that there was still no significant correlation between MRA findings and a PAR.

We reported sensitivity of MRA for fissuring of articular cartilage of 55% and 57% for articular cartilage thinning. These values are not dissimilar to those of Aprato et al. (2013), who investigated the accuracy of MRA in patients with a ‘radiological and clinical diagnosis of FAI’ using arthroscopy as the reference standard. They reported a sensitivity of 69% for acetabular cartilage lesions, and 46% and femoral head cartilage lesions. However, specificity for these lesions was higher in their study (88% and 81%) than our own (58% for thinning and 60% for fissuring).

The differences in accuracy of MRA in our study compared to those that have used arthroscopy as the reference standard can be interpreted in one of two ways. Firstly, it may be that our reference standard is not the gold standard and that findings based on surgical observations are more accurate. Alternatively, it may be that the tissue abnormalities identified during surgery are not necessarily symptomatic. Given the evidence that has demonstrated a high prevalence of pathological and/or structural

abnormalities in people who do not have any history of hip pain (Frank et al., 2015; Hack et al., 2010; Jung et al., 2011; Kang et al., 2010; Kumar et al., 2013; Lanyon et al., 2003; Lee et al., 2015; Register et al., 2012; Schmitz et al., 2012; Silvis et al., 2011), we are not convinced that surgical observation of pathology should be considered evidence that the source of pain has been identified.

## **7.5 Limitations**

A possible limitation of this study is that a single radiologist reported all MR images. It may be that a consensus of opinion from two or more radiologists would have improved the accuracy of the imaging findings, particularly given the evidence that has demonstrated that the intra-examiner reliability of radiologists in reading such images is only fair (Reurink et al., 2012). We were unable to get a second radiologist to agree to provide an independent report on all of the MR images. However, the radiologist in this study is a specialised musculoskeletal radiologist with 30 years of experience and is a recognised expert in this field. Previous research (McGuire et al., 2012) has demonstrated that such expertise is associated with a statistically significant higher level of accuracy. This level of expertise provides some assurance that the MRA findings reported in this study are an accurate interpretation of the MR images.

Another possible limitation is that whilst the radiologist was blinded to most clinical information about the participants, he was aware that they had been referred because of hip pain that might be of intra-articular origin. This knowledge may have created some bias that could have contributed to the relatively high prevalence of intra-articular pathology observed in this study. Whilst blinding the radiologist to all clinical information might have reduced any such bias, it would not ensure its removal. Given that participants were referred for an MRA and intra-articular injection of anaesthetic, it is obvious that referrer would suspect intra-articular pathology of the hip as that would be the only justification for a patient undergoing such a procedure. A case-control design would have enabled blinding of the radiologist however this type of study causes an over-estimation of diagnostic accuracy (Fritz & Wainner, 2001; Lijmer et al., 1999).

## 7.6 Conclusions

This study has demonstrated that a relatively high prevalence of abnormal findings is present in people with hip pain who are referred for an MRA. However, it has also demonstrated that the prevalence of abnormal findings is no different, apart for the presence of bone marrow oedema, between patients who have a positive response to an intra-articular injection of anaesthetic and those that don't. Our findings suggest that the presence or absence of such findings is not a key determinant for identification of symptomatic intra-articular pathology.

This study calls to question the appropriateness of making a diagnosis on the basis of MRA findings. It reinforces the growing body of evidence and opinion that imaging findings should only be used as an adjunct to those obtained from the history and clinical examination. The addition of a guided intra-articular injection of anaesthetic prior to MRA, along with assessment of the response to this anaesthetic by measuring the change in pain intensity reported during the application of provocative tests before and after the procedure, is recommended so that the relevance of pathology identified by MRA can be better determined.

## Chapter 8 Summary, Key Findings and Conclusions

This thesis contributes substantial new evidence that will inform the process of diagnosis and management of hip pain. This topic is important. Hip pain not only has the potential to have a significant effect on an individual's quality of life, it also places a significant cost burden on the health care system (Accident Compensation Corporation, 2015; Bennell, 2013). There is a consensus of opinion that the accurate identification of the cause of hip pain is difficult and that there is insufficient evidence to support the diagnostic utility of the clinical examination of the hip joint (Clohisy et al., 2009; McCarthy & Busconi, 1995; Reiman et al., 2013; Reiman et al., 2014b; Tijssen et al., 2012). Whilst a number of studies (Hase & Ueo, 1999; Keeney et al., 2004; Leunig, Werlen, Ungersböck, Ito, & Ganz, 1997; Narvani et al., 2003; Petersilge et al., 1996; Sink, Gralla, Ryba, & Dayton, 2008; Suenaga et al., 2002; Troelsen et al., 2009; Youdas et al., 2010) have investigated the relationship between some commonly used tests and various hip pathologies, the methodological quality of the majority of these studies is insufficient to enable definitive conclusions about the accuracy of the tests investigated (Reiman et al., 2013). Furthermore, this research was conducted primarily by orthopaedic specialists using arthroscopy as the reference standard. The included patients had hip pathology of sufficient severity to warrant orthopaedic assessment and surgical intervention. Any conclusions regarding the efficacy of the clinical examination in this cohort could not easily be generalised to the type of patients that would consult a physiotherapist in the primary health care setting.

One study (Sutlive et al., 2008), that has included participants from a primary care facility, used radiological findings as the reference standard for identifying hip osteoarthritis. These authors investigated the diagnostic accuracy of information collected from both the patient history and physical examination and reported that a number of variables had a statistically significant association ( $p \leq 0.10$ ) with positive radiographic findings. Sutlive and colleagues used logistic regression analysis to explore if various combinations of these variables might predict radiographic OA better than an individual variable. They reported a clinical prediction rule that included five physical tests (squat, hip flexion, scour, active extension and limited internal rotation ROM), stating that the presence of three out of five of these variables

increased the post-test probability of having radiological osteoarthritis to 68% (from a pre-test probability of 29%) and that four positive tests led to a 95% probability (positive LR of 24.3) of this finding. To date, Sutlive et al. are the only researchers that have developed a clinical prediction rule specifically for the purpose of identifying intra-articular hip pain. This rule has not subsequently been validated.

This study, in particular, provided the impetus for undertaking this doctoral thesis. It was clear that further investigation was required. Sutlive and colleagues had confined their investigation to the identification of osteoarthritis. Furthermore, the use of radiographs as the reference standard raised questions about the validity of their work considering what is known about the presence of abnormal findings on medical imaging in people with normal hips.

Hence, the current research was conceived, with the aim of providing additional and new evidence regarding the diagnostic accuracy of symptoms, signs and a wide variety of physical tests in a population of patients with hip pain that has not become severe enough to justify surgery. The thesis also explored novel methods of deriving a clinical prediction rule and a tool that enabled the assessment of the commonly recommended method of counting the number of positive tests to estimate the probability of the presence or absence of the condition of interest.

## **8.1 Key Findings**

### **Chapter 3: Pain Provocation and Pain Intensity During the Application of Physical Tests**

This study built upon current knowledge in three aspects of pain provocation testing that have a critical influence on how information collected from such tests is interpreted and utilised from a diagnostic perspective. Firstly, a common reason for employing physical tests during the patient examination is to determine which structure is the source of the patient's pain. Physical tests are presumed to load specific structures and to be provocative if the target structure is symptomatic. However, many normal structures will be painful if subjected to sufficient load. Prior to this study, little was known regarding the prevalence of painful responses to the application of hip tests in people with normal hips. This study has demonstrated that many commonly employed tests caused pain in *asymptomatic* hips. This finding suggests that a positive test does not necessarily implicate the presence of hip joint

pathology and adds weight to the existing evidence (Reiman et al., 2014a; Reiman et al., 2013) that demonstrates that physical examination tests of the hip are generally not specific.

The second aspect of pain provocation testing that this study addressed is closely related to the first. With the knowledge that such tests are provocative in many asymptomatic hips, it is important that *reproduction* of the patient's pain should be considered a positive test, not just the production of pain. Whilst three studies have examined the *inter-examiner* reliability of pain *production*, none of these studies required the patients to confirm that the pain produced was actually *the* pain that the patient felt when their hip was aggravated by day-to-day activities. No previous studies have investigated the *intra-examiner* reliability of the reproduction of a patient's pain with hip tests. Furthermore, although it is common for clinicians to ask patients to 'score' the intensity of pain reproduced with physical tests, little evidence to support the reliability of this information existed. The only previous study (Cliborne et al., 2004) that has examined this question performed hip tests on people with symptomatic *knee* OA, without indicating if a positive test was hip or knee pain, making it difficult interpret their results. Also, their study employed a test-retest period of only 2 minutes. Hence the need to assess the reliability of ratings of pain intensity, provoked by testing painful hips, over test-retest periods more relevant to clinical practice.

Thus, we investigated a large number of commonly employed pain provocation tests and demonstrated that most of these tests had 'moderate' to 'almost perfect' within-session (60 minutes apart) and between-session (2-7 days apart) reliability for both pain reproduction and ratings of pain intensity. These findings allow clinicians to be confident that patients can distinguish pain associated with their pathology (reproduced pain) from pain associated with loading normal structures (pain production) and, that patients can report pain intensity in a consistent manner.

Our results also demonstrated that the intensity of the pain provoked by testing had an influence on reliability. Generally, reliability improved as the level of pain experienced during testing increased. Tests that demonstrated poor reliability were primarily those that created pain of very low intensity. Our interpretation of these findings was that these *tests* should not be considered unreliable, but instead, *reports*

*of pain intensity* provoked by physical testing are unreliable *when* the pain intensity is low. Based on this evidence, we recommended that if assessment of pre to post treatment changes in pain intensity is required, the most provocative baseline tests should be employed. This should help ensure that reliable measures of pain intensity are obtained and therefore enable a more accurate assessment of any change in the patient's pain status. The degree of error that we observed with these measurements supports this recommendation. Based on our findings, tests that cause pain intensity of  $\geq 2$  points on the NPRS should be employed for such purposes.

#### **Chapter 4: Measurements of Strength and Range of Movement in Painful and Non-Painful Hips**

The diagnostic utility of measures of hip ROM and strength for identifying painful hip pathology has not been explored in depth. Whilst two previous studies (Arokoski et al., 2002; Rasch et al., 2007) have investigated side-to-side strength differences in people with unilateral hip pain, these studies only included patients with moderate to severe hip OA and reported some conflicting findings. With respect to ROM, a loss of hip internal rotation has been reported to be a predictor of hip OA (Altman et al., 1991; Birrell et al., 2001; Holla et al., 2012). However, two of these studies used radiographic evidence of OA as the reference standard and the third used a 'clinical diagnosis'. Consequently, none of these studies could confirm if this restriction in ROM was associated with pain arising from within the joint space. No studies have compared side-to-side differences in ROM from a diagnostic perspective.

Before the diagnostic accuracy of such measures could be investigated, the reliability of measurements of strength and ROM in people with unilateral hip pain needed to be established. Recent advances in technology have led to the development of new devices that have potential to be useful in both the research and clinical environments. This study considered the reliability of one such device that incorporates both a force transducer and gravity dependent inclinometer. The only previous studies (Pua et al., 2008; Sherrington & Lord, 2005) that have investigated the reliability of a hand-held force dynamometer in people with hip pain included relatively old patients (mean age 62 and 79 years respectively) with moderate to severe OA. Only one study (Pua et al.) has considered the reliability of a gravity inclinometer for measuring hip ROM in people with hip pain.

Thus, we measured ROM and strength of both hips in patients with unilateral hip pain. We purposely recruited younger patients (20-51 years) than those included in previous studies and patients with a range of pathologies (rather than just moderate to severe OA) so that our results were generalizable to a wider cohort of patients. Repeat measurements were made 60 minutes later and 2-7 days later. Reliability was determined and differences in strength and ROM between the symptomatic and asymptomatic hip were explored.

Our findings provided evidence of the reliability of this device and new knowledge in regard to the diagnostic utility of such measurements. With respect to peak isometric force in this group of patients with hip pain, excellent levels of reliability were observed (ICC values ranged from 0.70 to 0.92). Also, we provided information that quantified the percent error associated with these measurements, demonstrating a SEM% close to 10% for all muscle groups except adduction, where it was 16%. We did not observe any significant differences in strength between sides in our participants with unilateral hip pain. This was unexpected, as our patients had experienced pain for an average of just over 2 years. Although this indicates that the presence or absence of hip weakness does not appear to help to identify or exclude the presence of hip pathology, this study was powered to determine the reliability of tests and not side-to-side differences. Further research is necessary to confirm these findings

Our results also demonstrated that the gravity dependent inclinometer provided measures of ROM with high levels of reliability (ICCs ranging from 0.82 to 0.97). In contrast to strength measures, we observed a statistically significant reduction in range of movement between sides in our patients. However, this was only the case for the BKFO test and the mean difference was just 3.5 degrees. Although this is larger than the SEM ( $1.6^{\circ}$ ) associated with this measurement, the diagnostic utility of this finding may be limited given the size of this difference.

Despite the presence of pain and pathology, this study demonstrated that reliable measures of strength and ROM could be obtained with this relatively inexpensive, portable device. This finding provided the evidence required to justify the use of this device in the follow-up diagnostic accuracy study and supporting its use in the clinical setting, when objective assessment of changes in strength or ROM is required.

## **Chapter 5: The Diagnostic Accuracy of Findings from the Clinical Examination of the Painful Hip**

This study was designed to determine the diagnostic accuracy of symptoms and physical tests used in the clinical examination of the painful hip. Few previous studies have investigated the diagnostic accuracy of information obtained from the patient interview. Whilst a number of studies have investigated physical tests, the majority of these studies included patients that had undergone surgery, many were retrospective and some were case control studies. Also, across these studies, a limited number of tests were explored.

This study adds significantly to our current knowledge of hip diagnostics. The study was rigorously designed so that the risk of bias was very low. The findings have wide applicability due to the characteristics of the included participants and the reference standard employed. We purposely recruited people with hip pain that were likely to have less severe pathology than those included in studies that have employed arthroscopy as the reference standard. Although some authors consider arthroscopy as the *gold* standard, the inclusion criteria for these studies is narrow (i.e. the patient needs to be appropriate for surgery), making the findings hard to generalise to the large cohort patients who do not have hip pathology severe enough to require surgery. For similar reasons, we elected to use a FGAI as the reference standard. Patients do not need to have severe pathology to justify a FGAI. Therefore, we were able to investigate a broad range of patients, relevant to primary health care settings. Also, significant pain relief from an intra-articular FGAI is strong evidence that an intra-articular structure is the source of pain, whereas pathology observed surgically may not necessarily be symptomatic.

Our method of calculating the anaesthetic response to the FGAI was innovative. We compared the pain intensity scores reported by the patient *during the application* of provocative tests performed before the procedure, to that felt when these same tests were reapplied after the procedure. Previous studies (Martin et al., 2008; Maslowski et al., 2010) that have used this reference standard have asked the patients to score their pain pre-injection to that post-injection without actually requiring them to perform any provocative manoeuvres or undergoing reapplication of the tests being investigated. By repeating the provocative tests, we can be much more confident that

changes in pain intensity have been accurately assessed. To our knowledge, the current study is the first to use such a specific 'test-retest' requirement to evaluate changes in pain intensity.

With respect to the diagnostic utility of information obtained from the patient interview, the key findings of this study were the association between a positive anaesthetic response and the absence of dominant pain in the groin or the presence of crepitus. With respect to dominant pain in the groin, we reported a negative LR of 0.18 (95% CI 0.06, 0.58) and sensitivity of 0.91 (95% CI 0.77, 0.97). These values indicate that the absence of groin pain has utility as a screening test for intra-articular pathology of the hip. In contrast, the presence of crepitus had a high specificity (0.91 (95% CI 0.77, 0.97) and a moderate positive LR of 3.67 (95% CI 1.12 to 11.9). Based on these findings we recommended that these questions should be included in the clinical examination of the hip and given some weight in the diagnostic decision making process. Of the tests that we included in the physical examination, a number (FADDIR, FF, FFIR and quadrant) demonstrated high sensitivity. None of these tests had a negative likelihood ratio with confidence intervals that give assurance that a negative test significantly alters the probability of a PAR. However, the test with the best negative LR (0.14) was the quadrant. Hence, if a clinician wishes to screen for intra-articular pathology using a single physical test, we recommend that this test be employed but caution that false negatives cannot be ruled out. In regard to specificity, no physical test demonstrated sufficient diagnostic utility to indicate that it would be useful for identifying intra-articular pathology as stand-alone test.

Whilst previous studies have provided weak evidence of this nature, this study was the first to provide convincing evidence relevant to clinical practice.

### **Chapter 6: Predictors of Intra-articular Pathology of the Hip**

Building upon the findings of the previous chapter, a novel way of deriving clinical prediction rules and of utilising the findings of logistic regression analysis in the clinical setting was investigated. We also raised questions about the accuracy of a commonly reported method of combining multiple test results and proposed a new, sounder tool for aiding the diagnostic decision-making process. Clinical prediction rules are derived from information obtained from the patient examination. Typically this data is explored statistically to determine if there is any significant association

between any variable and the outcome of interest e.g. a specific diagnosis. The most common method of analysis is to enter (or remove) identified variables in a stepwise multiple regression model, based on the strength of this association, and to explore various combinations of variables to determine which combination of findings best predicts the diagnosis (Nathanson & Higgins, 2008). There does not appear to be any evidence that this method of identifying such predictors is the most appropriate way of doing so, despite limitations of the method being highlighted by statisticians (Harrell et al., 1996; Nathanson & Higgins, 2008). Similarly, there does not appear to be any published research in the fields of physiotherapy, musculoskeletal medicine or orthopaedics that uses any alternative method of developing a CPR. We looked critically at previous studies and considered alternative methods of analysis, choosing to employ information criterion as the means for selecting the optimal combinations of variables.

Hence, we conducted an in-depth, systematic exploration of all possible combinations of two or more key variables using a corrected version of the Akaike information criterion (AICc), and compared the findings of this analysis with that using the area under the curve (AUC) as the criterion. A significant advantage of the AICc is that it accounts for multicollinearity and reduces the probability of overfitting a model to essentially zero. The best overall model determined by this analysis included six predictors (dominant pain groin; age  $\geq 39$  years; the presence of crepitus; internal ROM  $< 41^\circ$ ; self-reported limited ROM and positive quadrant test). We provided detail of the regression co-efficient for each variable (test) in this model and a probability equation that allows estimation of the probability of an individual patient having a positive anaesthetic response based on the findings of the above tests. This model demonstrated an overall accuracy of 81%. Sensitivity and specificity were 91% and 70% respectively. The positive likelihood ratio was 3.1 and the negative LR was 0.12. Thus, the model has diagnostic utility for both ruling in and ruling out a PAR.

We proposed a simplified version of the probability equation that we called a 'Screening Score', and used a rescaled version of this score (the RSS) to assess the accuracy of the level of positivity (LOP) method of considering combinations of test results. This assessment highlighted that accuracy of the LOP approach to decision-making is *dependent* on there being relatively equal weightings in terms of the

contribution each test makes to the likelihood of a PAR. Where this is not the case, decisions made on the basis of the *number* of positive tests may be inaccurate. This finding may have important ramifications in other areas of medicine where combinations of test results, using variables identified by logistic regression, are employed using the LOP approach.

Our findings may have a significant impact on the clinical assessment and management of people with hip pain. If a positive anaesthetic response is an acceptable indication of the presence of intra-articular pathology, a means of accurately estimating the likelihood of such a response will greatly enhance the decision-making process. In this study, the post-test probability of a PAR increased to more than 80% with a RSS score of 4 or more. For a patient with this score, we recommend that referral for a guided intra-articular anaesthetic injection would be appropriate *if* surgical intervention was being considered. Conversely, a score of 3 or less would appear to be insufficient to justify an early referral for diagnostic injections, MR imaging or a surgical opinion. We believe that these guidelines may result in a more efficient use of resources and help to reduce the number of people undergoing unnecessary investigations and treatment. Similarly, they may decrease the delay in identifying intra-articular pathology that requires early recognition and specialist treatment.

### **Chapter 7: The Prevalence and Diagnostic Utility of Abnormal Findings Reported in Patients Undergoing Magnetic Resonance Imaging Arthrogram of the Hip**

Magnetic resonance imaging is widely utilised as a component of the clinical examination of the hip. However, the findings of such imaging need to be carefully considered given the evidence that a high prevalence of structural abnormalities has been reported in people without any history of hip pain (Frank et al., 2015; Kwee et al., 2013). Furthermore, there are concerns related to the reliability and accuracy of MR imaging (McGuire et al., 2012; Smith et al., 2011).

This study has made a unique contribution to our knowledge regarding the utility of MR imaging for identifying intra-articular pathology of the hip. We believe that it is the *first* study to determine the diagnostic accuracy of MR imaging of the hip using a fluoroscopy-guided anaesthetic injection of (FGAI) as the reference standard.

Previous studies have used arthroscopic findings as the reference standard. As previously discussed, arthroscopy is appropriate for a relatively small cohort of patients with hip pain, decreasing the generalisability of the findings of studies that have used this reference standard. Conversely, a FGAI is appropriate for a broader range of patients, and a significant reduction in pain (i.e. a positive anaesthetic response) after this procedure is strong evidence that intra-articular pathology is the source of the patient's symptoms (Bayer & Sekiya, 2010; Bogduk, 2004b; Martin et al., 2008).

In this study, MRA scans obtained for all participants in the preceding diagnostic accuracy study, were interpreted by an experienced musculoskeletal radiologist following a standardised format. The radiologist was blinded to all clinical information and from the response to a FGAI. Consistent with previous research, we observed a very high prevalence of abnormal findings, particularly of the acetabular labrum, but also of bony and cartilaginous tissue of the hip. Despite the high prevalence of structural abnormalities, the only finding that demonstrated a statistically significant association with a positive anaesthetic response was the presence of subchondral bone oedema. Notwithstanding this relationship, the sensitivity, specificity and likelihood ratios associated with this pathological finding demonstrated that it is unlikely to be clinically useful. Indeed, no individual structural abnormality demonstrated sufficient diagnostic accuracy to indicate that it has diagnostic utility as a stand-alone finding. These observations suggest that the presence of abnormalities identified by MRA should not be considered as evidence that an intra-articular source of hip pain has been established. Over-reliance on MR imaging as a diagnostic tool may lead to an increase in unnecessary surgical interventions.

On the basis of our findings, we suggested that MR imaging is not indicated unless there is a strong clinical suspicion of intra-articular pathology and possibly only when surgical intervention is being seriously contemplated. Considering our recommendation (in Chapter 6) that an RSS score of  $\geq 4$  should be required before an FGAI is performed, perhaps the appropriate level of suspicion for an MRA is when a patient has a RSS score of  $\geq 4$ , and has had a positive response to a FGAI.

## 8.2 Recommendations for future research

This research raised additional questions. We have identified the following areas of research:

1. Investigation of derived variables such as interactions through further logistic regression analysis using our existing data
2. Similarly, further exploration this data using alternative statistical analyses such as cluster analysis and the development of decision trees, will enable comparison to the findings obtained in the current study. If similar variables are identified as important through such analyses, this will provide additional support for the legitimacy of the current findings
3. Our clinical prediction rule needs further validation. A repeat study performed in a similar setting is required. Then, another in a different health care environment, with different examiners, referrers and radiologists.
4. These further diagnostic accuracy studies could also be extended so that a long-term follow-up of all included patients is included. Correlation between the clinical examinations, the FGAI and patient outcomes could then be assessed. In particular, the surgical findings of any participants that go on to have surgical intervention could be considered, as should the success or failure of this surgery.

## 8.3 Conclusion

The primary aims of of this thesis were to determine the diagnostic accuracy of information collected from the history and physical examination of the hip along with that obtained from magnetic resonance arthrogram. These aims were addressed through rigorously designed studies that examined the accuracy of individual variables to predict a positive anaesthetic response in patients who had hip joint pain.

This research provides robust evidence that demonstrates that the majority of physical tests are sensitive rather than specific. The test with the highest diagnostic utility was the quadrant test with accuracy values that demonstrate that a negative finding is useful in helping to screen for intra-articular pathology. This research also demonstrated the importance of information obtained from the patient history with the absence of dominant pain in the groin proving useful for ruling out symptomatic intra-

articular pathology and the presence of crepitus being useful for identifying this pathology. The value of the patient history was reflected in the clinical prediction rule derived from the data collected from the clinical examination. Four of the six variables identified were obtained from the history i.e. crepitus, dominant pain in the groin, age >39 and self-reported limitation of ROM. The only physical findings that were included were a positive quadrant test and internal rotation ROM less than 41°.

This thesis proposed a screening score as novel method of utilising the information contained within clinical prediction rules. This score appropriately weights the contribution that each of the variables in the CPR make towards the diagnosis of intra-articular pain. This has both research and clinical value. In particular, the CPR and associated screening score have the potential to significantly enhance decision-making for patients with pain in the hip region. Our results suggest that it may not be justified to refer patients with a screening score of 3 or less for diagnostic injections, MR imaging or a surgical opinion in the initial stages of their management. This approach may result in a reduction in both the number of unnecessary referrals for medical imaging and the number of surgical interventions undertaken as a result of abnormal pathology being identified with such imaging. Our findings regarding the poor diagnostic accuracy of abnormal morphology identified by MRA reinforces the importance of considering imaging findings in the context of the clinical examination.

The clinical prediction rule should be considered preliminary until it has been validated in follow-up studies. However, given the potential to decrease the costs associated with diagnostic imaging and perhaps unnecessary escalation to diagnostic or interventional arthroscopy, we believe that it would be appropriate for clinicians managing patients with hip pain to consider applying the findings of this research prior to such validation.

## References

- Abe, I., Harada, Y., Oinuma, K., Kamikawa, K., Kitahara, H., Morita, F., & Moriya, H. (2000). Acetabular labrum: abnormal findings at MR imaging in asymptomatic hips. *Radiology*, *216*(2), 576-581. doi:10.1148/radiology.216.2.r00au13576
- Access Economics. (2010). *The economic cost of arthritis in New Zealand in 2010, Report for Arthritis New Zealand*. Canberra. Retrieved from <http://www.arthritis.org.nz/wp-content/uploads/2011/07/economic-cost-of-arthritis-in-new-zealand-final-print.pdf>
- Accident Compensation Corporation. (2015). ACC Injury Statistics Tool. Retrieved 25 February 2015 <http://www.acc.co.nz/about-acc/statistics/injury-statistics/index.htm>
- Agricola, R., Heijboer, M. P., Bierma-Zeinstra, S. M. A., Verhaar, J. A. N., Weinans, H., & Waarsing, J. H. (2013). Cam impingement causes osteoarthritis of the hip: a nationwide prospective cohort study (CHECK). *Annals of the Rheumatic Diseases*, *72*(6), 918-923. doi:10.1136/annrheumdis-2012-201643
- Aliu, O., & Chung, K. C. (2012). Assessing strength of evidence in diagnostic tests. *Plastic and Reconstructive Surgery*, *129*(6), 989e-998e. doi:10.1097/PRS.0b013e31824ecd61
- Altman, R., Alarcon, G., Appelrouth, D., Bloch, D., Borenstein, D., Brandt, K., . . . McDonald, E. (1991). The American college of rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis and Rheumatism*, *34*(5), 505-514.
- Altman, R., & Gold, G. (2007). Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and Cartilage*, *15*(SUPPL. 1), 1-56.
- Aprato, A., Massè, A., Faletti, C., Valente, A., Atzori, F., Stratta, M., & Jayasekera, N. (2013). Magnetic resonance arthrography for femoroacetabular impingement surgery: is it reliable? *Journal of Orthopaedics and Traumatology*, *14*(3), 201-206. doi:10.1007/s10195-013-0227-1
- Arnold, C. M., Warkentin, K. D., Chilibeck, P. D., & Magnus, C. R. A. (2010). The reliability and validity of handheld dynamometry for the measurement of lower-extremity muscle strength in older adults. *Journal of Strength and Conditioning Research*, *24*(3), 815-824.
- Arnold, D. R., Keene, J. S., Blankenbaker, D. G., & DeSmet, A. A. (2011). Hip pain referral patterns in patients with labral tears: analysis based on intra-articular anesthetic injections, hip arthroscopy, and a new pain "circle" diagram. *Physician and Sportsmedicine*, *39*(1), 29-35. doi:10.3810/psm.2011.02.1839
- Arokoski, M. H., Arokoski, J. P. A., Haara, M., Kankaanpaa, M., Vesterinen, M., Niemitukia, L. H., & Helminen, H. J. (2002). Hip muscle strength and muscle cross sectional area in men with and without hip osteoarthritis. *Journal of Rheumatology*, *29*(10), 2185-2195.
- Arokoski, M. H., Haara, M., Helminen, H. J., & Arokoski, J. P. (2004). Physical function in men with and without hip osteoarthritis. *Archives of Physical Medicine and Rehabilitation*, *85*(4), 574-581. doi:10.1016/j.apmr.2003.07.011
- Ashok, N., Sivan, M., Tafazal, S., & Sell, P. (2009). The diagnostic value of anaesthetic hip injection in differentiating between hip and spinal pain. *European Journal of Orthopaedic Surgery & Traumatology*, *19*(3), 167-171.

- Assouline-Dayana, Y., Chang, C., Greenspan, A., Shoenfeld, Y., & Gershwin, M. E. (2002). Pathogenesis and natural history of osteonecrosis. *Seminars in Arthritis and Rheumatism*, 32(2), 94-124. doi:10.1053/sarh.2002.33724b
- Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26(4), 217-238. doi:10.2165/00007256-199826040-00002
- Ayeni, O., Chu, R., Hetaimish, B., Nur, L., Simunovic, N., Farrokhyar, F., . . . Bhandari, M. (2014a). A painful squat test provides limited diagnostic utility in CAM-type femoroacetabular impingement. *Knee Surgery, Sports Traumatology, Arthroscopy*, 22(4), 806-811. doi:10.1007/s00167-013-2668-8
- Ayeni, O. R., Farrokhyar, F., Crouch, S., Chan, K., Sprague, S., & Bhandari, M. (2014b). Pre-operative intra-articular hip injection as a predictor of short-term outcome following arthroscopic management of femoroacetabular impingement. *Knee Surgery, Sports Traumatology, Arthroscopy*, 22(4), 801-805. doi:10.1007/s00167-014-2883-y
- Bachmann, L. M., Puhan, M. A., Riet, G. t., & Bossuyt, P. M. (2006). Sample sizes of studies on diagnostic accuracy: literature survey [Journal Article]. *British Medical Journal*, 332, 1127-1129. doi:10.1136/bmj.38793.637789.2F
- Bandinelli, S., Benvenuti, E., Del Lungo, I., Baccini, M., Benvenuti, F., Di Iorio, A., & Ferrucci, L. (1999). Measuring muscular strength of the lower limbs by hand-held dynamometer: a standard protocol. *Aging Clinical and Experimental Research*, 11(5), 287-293.
- Baron, R., Binder, A., & Wasner, G. (2010). Neuropathic pain: diagnosis, pathophysiological mechanisms, and treatment. *The Lancet Neurology*, 9(8), 807-819. doi:[http://dx.doi.org/10.1016/S1474-4422\(10\)70143-5](http://dx.doi.org/10.1016/S1474-4422(10)70143-5)
- Bayer, J. L., & Sekiya, J. K. (2010). Hip instability and capsular laxity. *Operative Techniques in Orthopaedics*, 20(4), 237-241. doi:<http://dx.doi.org/10.1053/j.oto.2010.09.019>
- Beattie, P., & Nelson, R. (2006). Clinical prediction rules: what are they and what do they tell us? *Australian Journal of Physiotherapy*, 52(3), 157-163.
- Bennell, K. (2013). Physiotherapy management of hip osteoarthritis. *Journal of Physiotherapy*, 59(3), 145-157. doi:[http://dx.doi.org/10.1016/S1836-9553\(13\)70179-6](http://dx.doi.org/10.1016/S1836-9553(13)70179-6)
- Bennett, M. I., Smith, B. H., Torrance, N., & Potter, J. (2005). The S-LANSS score for identifying pain of predominantly neuropathic origin: validation for use in clinical and postal research. *The Journal of Pain*, 6(3), 149-158.
- Bergmann, G., Graichen, F., & Rohlmann, A. (2004). Hip joint contact forces during stumbling. *Langenbeck's Archives of Surgery*, 389(1), 53-59. doi:10.1007/s00423-003-0434-y
- Bierma-Zeinstra, S. M. A., Bohnen, A. M., Ramlal, R., Ridderikhoff, J., Verhaar, J. A. N., & Prins, A. (1998). Comparison between two devices for measuring hip joint motions. *Clinical Rehabilitation*, 12(6), 497-505.
- Bijlsma, J. W. J., Berenbaum, F., & Lafeber, F. P. J. G. (2011). Osteoarthritis: an update with relevance for clinical practice. *The Lancet*, 377(9783), 2115-2126. doi:[http://dx.doi.org/10.1016/S0140-6736\(11\)60243-2](http://dx.doi.org/10.1016/S0140-6736(11)60243-2)
- Birrell, F., Croft, P., Cooper, C., Hosie, G., Macfarlane, G., & Silman, A. (2001). Predicting radiographic hip osteoarthritis from range of movement. *Rheumatology*, 40(5), 506-512.
- Birrell, F., Lunt, M., Macfarlane, G., & Silman, A. (2005). Association between pain in the hip region and radiographic changes of osteoarthritis: results from a

- population-based study. *Rheumatology*, 44(3), 337-341.  
doi:10.1093/rheumatology/keh458
- Blankenbaker, D. G., De Smet, A. A., Keene, J. S., & Fine, J. P. (2007). Classification and localization of acetabular labral tears. *Skeletal Radiology*, 36(5), 391-397.
- Bloom, N., & Cornbleet, S. L. (2014). Hip rotator strength in healthy young adults measured in hip flexion and extension by using a hand-held dynamometer. *Physical Medicine and Rehabilitation*, 6(12), 1137-1142.  
doi:10.1016/j.pmrj.2014.06.002
- Bogduk, N. (2004a). Diagnostic blocks: a truth serum for malingering. *Clinical Journal of Pain*, 20(6), 409-414. doi:10.1097/00002508-200411000-00005
- Bogduk, N. (2004b). *Practice guidelines for spinal diagnostic and treatment procedures* (1st ed.). San Francisco: International Spine Intervention Society.
- Bogduk, N. (2009). On the definitions and physiology of back pain, referred pain, and radicular pain. *Pain*, 147(1-3), 17-19.
- Bohannon, R. W. (1986). Test-retest reliability of hand-held dynamometry during a single session of strength assessment. *Physical Therapy*, 66(2), 206-208.
- Bohannon, R. W. (1988). Make tests and break tests of elbow flexor muscle strength. *Physical Therapy*, 68(2), 193-194.
- Bohannon, R. W. (2012). Hand-held dynamometry: A practicable alternative for obtaining objective measures of muscle strength. *Isokinetics and Exercise Science*, 20(4), 301-315.
- Bohannon, R. W., & Saunders, N. (1990). Hand-held dynamometry: a single trial may be adequate for measuring muscle strength in healthy individuals. *Physiotherapy Canada*, 42(1), 6-9.
- Bohannon, R. W., Vigneault, J., & Rizzo, J. (2008). Hip external and internal rotation strength: consistency over time and between sides. *Isokinetics and Exercise Science*, 16(2), 107-111.
- Bonsell, S., Pearsall IV, A. W., Heitman, R. J., Helms, C. A., Major, N. M., & Speer, K. P. (2000). The relationship of age, gender, and degenerative changes observed on radiographs of the shoulder in asymptomatic individuals. *Journal of Bone and Joint Surgery*, 82(8), 1135-1139.
- Boone, D. C., & Azen, S. P. (1979). Normal range of motion of joints in male subjects. *Journal of Bone and Joint Surgery - Series A*, 61(5), 756-759.
- Boos, N., Semmer, N., Elfering, A., Schade, V., Gal, I., Zanetti, M., . . . Main, C. J. (2000). Natural history of individuals with asymptomatic disc abnormalities in magnetic resonance imaging. *Spine*, 25(12), 1484-1492.  
doi:10.1097/00007632-200006150-00006
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de vet, H. C. W. (2003a). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Annals of Internal Medicine*, 138(1), 40.
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . Lijmer, J. G. (2003b). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*, 138(1), 1.
- Botser, I. B., Martin, D. E., Stout, C. E., & Domb, B. G. (2011). Tears of the ligamentum teres: prevalence in hip arthroscopy using 2 classification systems. *American Journal of Sports Medicine*, 39, 117S-125S.

- Brandt-Rauf, P. W., & Brandt-Rauf, S. I. (1987). History of occupational medicine: relevance of Imhotep and the Edwin Smith papyrus. *British Journal of Industrial Medicine*, 44(1), 68-70.
- Brinjikji, W., Luetmer, P. H., Comstock, B., Bresnahan, B. W., Chen, L. E., Deyo, R. A., . . . Jarvik, J. G. (2015). Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *AJNR: American Journal of Neuroradiology*, 36(4), 811-816. doi:10.3174/ajnr.A4173
- Brown, L. E., & Weir, J. P. (2001). ASEP procedures recommendation I: accurate assessment of muscular strength and power. *Journal of Exercise Physiology Online*, 4(3), 1-21.
- Burgess, R. M., Rushton, A., Wright, C., & Daborn, C. (2011). The validity and accuracy of clinical diagnostic tests used to detect labral pathology of the hip: a systematic review. *Manual Therapy*, 16(4), 318-326. doi:<http://dx.doi.org/10.1016/j.math.2011.01.002>
- Burnett, R. S. J., Della Rocca, G. J., Prather, H., Curry, M., Maloney, W. J., & Clohisy, J. C. (2006). Clinical presentation of patients with tears of the acetabular labrum. *Journal of Bone and Joint Surgery (American Volume)*, 88(7), 1448-1457.
- Byrd, J. W. T. (2014). Femoroacetabular impingement in athletes: current concepts. *American Journal of Sports Medicine*, 42(3), 737-751. doi:10.1177/0363546513499136
- Byrd, J. W. T., & Jones, K. S. (2001). Hip arthroscopy in athletes. *Clinics in Sports Medicine*, 20(4), 749-762.
- Byrd, J. W. T., & Jones, K. S. (2004a). Diagnostic accuracy of clinical assessment, magnetic resonance imaging, magnetic resonance arthrography, intra-articular injection in hip arthroscopy patients. *American Journal of Sports Medicine*, 32(7), 1668-1674.
- Byrd, J. W. T., & Jones, K. S. (2004b). Traumatic rupture of the ligamentum teres as a source of hip pain. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 20(4), 385-391. doi:<http://dx.doi.org/10.1016/j.arthro.2004.01.025>
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429.
- Cadogan, A., Laslett, M., Hing, W., McNair, P., & Williams, M. (2010). Interexaminer reliability of orthopaedic special tests used in the assessment of shoulder pain. *Manual Therapy*.
- Cadogan, A., Laslett, M., Hing, W., McNair, P., & Williams, M. (2011). Interexaminer reliability of orthopaedic special tests used in the assessment of shoulder pain. *Manual Therapy*, 16(2), 131-135. doi:10.1016/j.math.2010.07.009
- Cadogan, A., McNair, P., Laslett, M., & Hing, W. (2013). Shoulder pain in primary care: diagnostic accuracy of clinical examination tests for non-traumatic acromioclavicular joint pain. *BMC Musculoskeletal Disorders*, 14. doi:10.1186/1471-2474-14-156
- Casartelli, N. C., Maffiuletti, N. A., Item-Glatthorn, J. F., Staehli, S., Bizzini, M., Impellizzeri, F. M., & Leunig, M. (2011). Hip muscle weakness in patients with symptomatic femoroacetabular impingement. *Osteoarthritis and Cartilage*, 19(7), 816-821.
- Cerezal, L., Kassarian, A., Canga, A., Dobado, M. C., Montero, J. A., Llopis, E., . . . Perez-Carro, L. (2010). Anatomy, biomechanics, imaging, and management of ligamentum teres injuries. *Radiographics*, 30(6), 1637-1651.

- Childs, J. D., & Cleland, J. A. (2006). Development and application of clinical prediction rules to improve decision making in physical therapist practice. *Physical Therapy, 86*(1), 122-131.
- Childs, J. D., Piva, S. R., & Fritz, J. M. (2005). Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine, 30*(11), 1331-1334.
- Chinn, S. (1991). Statistics in respiratory medicine 2. Repeatability and method comparison. *Thorax, 46*(6), 454-456.
- Chong, T., Don, D. W., Kao, M., Wong, D., & Mitra, R. (2013). The value of physical examination in the diagnosis of hip osteoarthritis. *Journal of Back and Musculoskeletal Rehabilitation, 26*(4), 397-400.
- Chronopoulos, E., Kim, T. K., Park, H. B., Ashenbrenner, D., & McFarland, E. G. (2004). Diagnostic value of physical tests for isolated chronic acromioclavicular lesions. *American Journal of Sports Medicine, 32*(3), 655-661. doi:10.1177/0363546503261723
- Cibere, J., Thorne, A., Bellamy, N., Greidanus, N., Chalmers, A., Mahomed, N., . . . Esdaile, J. M. (2008). Reliability of the hip examination in osteoarthritis: effect of standardization. *Arthritis Care and Research, 59*(3), 373-381.
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551-558. doi:10.1016/0895-4356(90)90159-M
- Cichanowski, H. R., Schmitt, J. S., Johnson, R. J., & Niemuth, P. E. (2007). Hip strength in collegiate female athletes with patellofemoral pain. *Medicine and Science in Sports and Exercise, 39*(8), 1227-1232. doi:10.1249/mss.0b013e3180601109
- Cleland, J. A., Childs, J. D., Fritz, J. M., Whitman, J. M., & Eberhart, S. L. (2007). Development of a clinical prediction rule for guiding treatment of a subgroup of patients with neck pain: use of thoracic spine manipulation, exercise, and patient education. *Physical Therapy, 87*(1), 9-23. doi:10.2522/ptj.20060155
- Cleland, J. A., Mintken, P. E., Carpenter, K., Fritz, J. M., Glynn, P., Whitman, J., & Childs, J. D. (2010). Examination of a clinical prediction rule to identify patients with neck pain likely to benefit from thoracic spine thrust manipulation and a general cervical range of motion exercise: multi-center randomized clinical trial. *Physical Therapy, 90*(9), 1239-1250. doi:10.2522/ptj.20100123
- Cliborne, A. V., Wainner, R. S., Rhon, D. I., Judd, C. D., Fee, T. T., Matekel, R. L., & Whitman, J. M. (2004). Clinical hip tests and a functional squat test in patients with knee osteoarthritis: reliability, prevalence of positive test findings, and short-term response to hip mobilization. *Journal of Orthopaedic and Sports Physical Therapy, 34*(11), 676-685. doi:10.2519/jospt.2004.1432
- Clohisey, J. C., Knaus, E. R., Hunt, D. M., Leshner, J. M., Harris-Hayes, M., & Prather, H. (2009). Clinical presentation of patients with symptomatic anterior hip impingement. *Clinical Orthopaedics and Related Research, 467*(3), 638-644.
- Coldham, F., Lewis, J., & Lee, H. (2006). The reliability of one vs. three grip trials in symptomatic and asymptomatic subjects. *Journal of Hand Therapy, 19*(3), 318-327. doi:10.1197/j.jht.2006.04.002
- Collins, J. A., Ward, J. P., & Youm, T. (2014). Is prophylactic surgery for femoroacetabular impingement indicated?: a systematic review. *American Journal of Sports Medicine, 42*(12), 3009-3015. doi:10.1177/0363546513499227

- Connor, P. M., Banks, D. M., Tyson, A. B., Coumas, J. S., & D'Alessandro, D. F. (2003). Magnetic resonance imaging of the asymptomatic shoulder of overhead athletes: a 5-year follow-up study. *American Journal of Sports Medicine*, *31*(5), 724-727.
- Cook, C. (2010). The lost art of the clinical examination: An overemphasis on clinical special tests. *Journal of Manual & Manipulative Therapy*, *18*, 3-4.
- Cook, C., Brown, C., Isaacs, R., Roman, M., Davis, S., & Richardson, W. (2010). Clustered clinical findings for diagnosis of cervical spine myelopathy. *Journal of Manual and Manipulative Therapy*, *18*(4), 175-180. doi:10.1179/106698110x12804993427045
- Coomes, E. N. (1963). Experimental pain from the hip-joint. *Annals of Physical Medicine*, *201*, 100-104.
- Courtney, C. A., Kavchak, A. E., Lowry, C. D., & O'Hearn, M. A. (2010). Interpreting joint pain: quantitative sensory testing in musculoskeletal management. *Journal of Orthopaedic and Sports Physical Therapy*, *40*(12), 818-825. doi:10.2519/jospt.2010.3314
- Cox, D. (1970). *The analysis of binary data*. London: Methuen.
- Croft, P., Cooper, C., Wickham, C., & Coggon, D. (1990). Defining osteoarthritis of the hip for epidemiologic studies. *American Journal of Epidemiology*, *132*(3), 514-522.
- Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, *84*(8), 1022-1028.
- Croskerry, P., Singhal, G., & Mamede, S. (2013a). Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Quality and Safety*, *22*(SUPPL.2), ii58-ii64.
- Croskerry, P., Singhal, G., & Mamede, S. (2013b). Cognitive debiasing 2: impediments to and strategies for change. *BMJ Quality and Safety*, *22*(SUPPL.2), ii65-ii72. doi:10.1136/bmjqs-2012-001712
- Cyriax, J. (1974). *Textbook of Orthopaedic Medicine; Volume One: Diagnosis of Soft Tissue Lesions*. London: Bailliere Tindall.
- Czerny, C., Hofmann, S., Urban, M., Tschauer, C., Neuhold, A., Pretterklieber, M., . . . Kramer, J. (1999). MR arthrography of the adult acetabular capsular-labral complex: correlation with surgery and anatomy. *American Journal of Roentgenology*, *173*(2), 345-349.
- Datir, A., Xing, M., Kang, J., Harkey, P., Kakarala, A., Carpenter, W. A., & Terk, M. R. (2014). Diagnostic utility of mri and mr arthrography for detection of ligamentum teres tears: a retrospective analysis of 187 patients with hip pain. *American Journal of Roentgenology*, *203*(2), 418-423. doi:10.2214/ajr.13.12258
- Davidson, M. (2002). The interpretation of diagnostic tests: a primer for physiotherapists. *Australian Journal of Physiotherapy*, *48*(3), 227-232.
- Delitto, A., & Snyder-Mackler, L. (1995). The diagnostic process: examples in orthopedic physical therapy. *Physical Therapy*, *75*(3), 203-211.
- Denegar, C. R., & Fraser, M. (2006). How useful are physical examination procedures? Understanding and applying likelihood ratios. *Journal of Athletic Training*, *41*(2), 201-206.
- Deshmukh, A. J., Thakur, R. R., Goyal, A., Klein, D. A., Ranawat, A. S., & Rodriguez, J. A. (2010). Accuracy of diagnostic injection in differentiating source of atypical hip pain. *Journal of Arthroplasty*. doi:10.1016/j.arth.2010.04.015

- Devitt, B. M., Philippon, M. J., Goljan, P., Peixoto, L. P., Briggs, K. K., & Ho, C. P. (2014). Preoperative diagnosis of pathologic conditions of the ligamentum teres: is mri a valuable imaging modality? *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, *30*(5), 568-574.
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., . . . Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155-164. doi:10.1111/j.1547-5069.2007.00161.x
- Deyo, R. A. (2013). Real help and red herrings in spinal imaging. *The New England Journal of Medicine*, *368*(11), 1056-1058.
- Diamond, L. E., Dobson, F. L., Bennell, K. L., Wrigley, T. V., Hodges, P. W., & Hinman, R. S. (2015). Physical impairments and activity limitations in people with femoroacetabular impingement: a systematic review. *British Journal of Sports Medicine*, *49*(4), 230-242. doi:10.1136/bjsports-2013-093340
- Dobson, F., Choi, Y. M., Hall, M., & Hinman, R. S. (2012). Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: a systematic review. *Arthritis Care and Research*, *64*(10), 1565-1575.
- Domb, B. G., Brooks, A. G., & Byrd, J. W. T. (2009). Clinical examination of the hip joint in athletes. *Journal of Sport Rehabilitation*, *18*(1), 3-23.
- Dorleijn, D. M. J., Luijsterburg, P. A. J., Bierma-Zeinstra, S. M. A., & Bos, P. K. (2014). Is anesthetic hip joint injection useful in diagnosing hip osteoarthritis? A meta-analysis of case series. *Journal of Arthroplasty*, *29*(6), 1236-1242.e1231. doi:<http://dx.doi.org/10.1016/j.arth.2013.12.008>
- Dreyfuss, P., Michaelsen, M., Pauza, K., McLarty, J., & Bogduk, N. (1996). The value of medical history and physical examination in diagnosing sacroiliac joint pain. *Spine*, *21*(22), 2594-2602.
- Dunn, G. (1992). Design and analysis of reliability studies. *Statistical Methods in Medical Research*, *1*(2), 123-157.
- Elstein, A. S. (2009). Thinking about diagnostic thinking: a 30-year perspective. *Advances in Health Sciences Education*, *14*(1 SUPPL), 7-18. doi:10.1007/s10459-009-9184-0
- Elstein, A. S., & Schwarz, A. (2002). Evidence base of clinical diagnosis. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal*, *324*(7339), 729-732.
- Eng, J. (2005). Receiver operating characteristic analysis: a primer. *Academic Radiology*, *12*(7), 909-916.
- Fagan, T. J. (1975). Letter: Nomogram for Bayes theorem. *The New England Journal of Medicine*, *293*(5), 257.
- Farrar, J. T., Young, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*, *94*(2), 149-158.
- Feddock, C. A. (2007). The lost art of clinical skills. *American Journal of Medicine*, *120*(4), 374-378.
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549. doi:10.1016/0895-4356(90)90158-1
- Fenn Buderer, N. M. (1996). Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine*, *3*(9), 895-900.

- Fernández-De-Las-Peñas, C., Cleland, J. A., Cuadrado, M. L., & Pareja, J. A. (2008). Predictor variables for identifying patients with chronic tension-type headache who are likely to achieve short-term success with muscle trigger point therapy. *Cephalalgia*, 28(3), 264-275. doi:10.1111/j.1468-2982.2007.01530.x
- Flynn, T., Fritz, J., Whitman, J., Wainner, R., Magel, J., Rendeiro, D., . . . Allison, S. (2002). A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine*, 27(24), 2835-2843.
- Frank, J. M., Harris, J. D., Erickson, B. J., Slikker, W., III, Bush-Joseph, C. A., Salata, M. J., & Nho, S. J. (2015). Prevalence of femoroacetabular impingement imaging findings in asymptomatic volunteers: a systematic review. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 31(6), 1199-1204. doi:10.1016/j.arthro.2014.11.042
- Freehill, M. T., & Safran, M. R. (2011). The labrum of the hip: diagnosis and rationale for surgical correction. *Clinics in Sports Medicine*, 30(2), 293-315. doi:10.1016/j.csm.2010.12.002
- Fritz, J. M. (2009). Clinical prediction rules in physical therapy: coming of age? *Journal of Orthopaedic and Sports Physical Therapy*, 39(3), 159-161.
- Fritz, J. M., Piva, S. R., & Childs, J. D. (2005). Accuracy of the clinical examination to predict radiographic instability of the lumbar spine. *European Spine Journal*, 14(8), 743-750.
- Fritz, J. M., & Wainner, R. S. (2001). Examining diagnostic tests: an evidence-based perspective. *Physical Therapy*, 81(9), 1546-1564.
- Fulcher, M. L., Hanna, C. M., & Elley, C. R. (2010). Reliability of handheld dynamometry in assessment of hip strength in adult male football players. *Journal of Science and Medicine in Sport*, 13(1), 80-84.
- Gajdosik, R. L., & Bohannon, R. W. (1987). Clinical measurement of range of motion. Review of goniometry emphasizing reliability and validity. *Physical Therapy*, 67(12), 1867-1872.
- Ganz, R., Leunig, M., Leunig-Ganz, K., & Harris, W. H. (2008). The etiology of osteoarthritis of the hip: an integrated mechanical concept. *Clinical Orthopaedics and Related Research*, 466(2), 264-272.
- Gerhardt, M., Johnson, K., Atkinson, R., Snow, B., Shaw, C., Brown, A., & Thomas Vangsness Jr, C. (2012). Characterisation and classification of the neural anatomy in the human hip joint. *Hip International*, 22(1), 75-81. doi:10.5301/hip.2012.9042
- Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11), 1129-1135. doi:10.1016/s0895-4356(03)00177-x
- Gordon, A. M., Huxley, A. F., & Julian, F. J. (1966). The variation in isometric tension with sarcomere length in vertebrate muscle fibres. *Journal of Physiology*, 184(1), 170-192.
- Grajski, K. A., Breiman, L., Di Prisco, G. V., & Freeman, W. J. (1986). Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART. *IEEE Transactions on Biomedical Engineering*, BME-33(12), 1076-1086. doi:10.1109/TBME.1986.325684
- Grant, A. D., Sala, D. A., & Davidovitch, R. I. (2012). The labrum: structure, function, and injury with femoro-acetabular impingement. *Journal of Children's Orthopaedics*, 6(5), 357-372. doi:10.1007/s11832-012-0431-1

- Grassel, S. (2014). The role of peripheral nerve fibers and their neurotransmitters in cartilage and bone physiology and pathophysiology. *Arthritis Research and Therapy*, 16(1). doi:10.1186/s13075-014-0485-1
- Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, 79(3), 340-349.
- Griner, P. F., Mayewski, R. J., Mushlin, A. I., & Greenland, P. (1981). Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Annals of Internal Medicine*, 94(4 Pt 2), 557-592.
- Groh, M. M., & Herrera, J. (2009). A comprehensive review of hip labral tears. *Current Reviews in Musculoskeletal Medicine*, 2(2), 105-117. doi:10.1007/s12178-009-9052-9
- Gruther, W., Wick, F., Paul, B., Leitner, C., Posch, M., Matzner, M., . . . Ebenbichler, G. (2009). Diagnostic accuracy and reliability of muscle strength and endurance measurements in patients with chronic low back pain [Article]. *Journal of Rehabilitation Medicine*, 41(8), 613-619. doi:10.2340/16501977-0391
- Guyatt, G., Walter, S., Shannon, H., Cook, D., Jaeschke, R., & Heddle, N. (1995). Basic statistics for clinicians: 4. Correlation and regression. *Canadian Medical Association Journal*, 152(4), 497-504.
- Gwilym, S. E., Filippini, N., Douaud, G., Carr, A. J., & Tracey, I. (2010). Thalamic atrophy associated with painful osteoarthritis of the hip is reversible after arthroplasty: a longitudinal voxel-based morphometric study. *Arthritis and Rheumatism*, 62(10), 2930-2940. doi:10.1002/art.27585
- Gwilym, S. E., Keltner, J. R., Warnaby, C. E., Carr, A. J., Chizh, B., Chessell, I., & Tracey, I. (2009). Psychophysical and functional imaging evidence supporting the presence of central sensitization in a cohort of osteoarthritis patients. *Arthritis Care and Research*, 61(9), 1226-1234. doi:10.1002/art.24837
- Haas, M. (1991). Statistical methodology for reliability studies. *Journal of Manipulative and Physiological Therapeutics*, 14(2), 119-132.
- Hack, K., Di Primio, G., Rakhra, K., & Beaulé, P. E. (2010). Prevalence of cam-type femoroacetabular impingement morphology in asymptomatic volunteers. *Journal of Bone and Joint Surgery - Series A*, 92(14), 2436-2444. doi:10.2106/JBJS.J.01280
- Hagen, M. D. (1995). Test characteristics: how good is that test? *Primary Care: Clinics in Office Practice*, 22(2), 213-233.
- Hampton, J. R., Harrison, M. J. G., Mitchell, J. R. A., Prichard, J. S., & Seymour, C. (1975). Relative contributions of history taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *British Medical Journal*, 2(5969), 486-489.
- Hananouchi, T., Yasui, Y., Yamamoto, K., Toritsuka, Y., & Ohzono, K. (2012). Anterior impingement test for labral lesions has high positive predictive value. *Clinical Orthopaedics and Related Research*, 470(12), 3524-3529. doi:10.1007/s11999-012-2450-0
- Haneline, M. T. (2007). A review of the use of likelihood ratios in the chiropractic literature. *Journal of Chiropractic Medicine*, 6(3), 99-104.
- Harper, D. (2015). *Online Etymology Dictionary*. Retrieved 26 June, 2015, from <http://www.etymonline.com>
- Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and

- measuring and reducing errors. *Statistics in Medicine*, 15(4), 361-387.  
doi:10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4
- Harris, J. D., McCormick, F. M., Abrams, G. D., Gupta, A. K., Ellis, T. J., Bach Jr, B. R., . . . Nho, S. J. (2013). Complications and reoperations during and after hip arthroscopy: a systematic review of 92 studies and more than 6,000 patients. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 29(3), 589-595. doi:<http://dx.doi.org/10.1016/j.arthro.2012.11.003>
- Harris-Hayes, M., Mueller, M. J., Sahrman, S. A., Bloom, N. J., Steger-May, K., Clohisy, J. C., & Salsich, G. B. (2014). Persons with chronic hip joint pain exhibit reduced hip muscle strength. *Journal of Orthopaedic and Sports Physical Therapy*, 44(11), 890-898. doi:10.2519/jospt.2014.5268
- Hart, D. L., Stobbe, T. J., Till, C. W., & Plummer, R. W. (1984). Effect of trunk stabilization on quadriceps femoris muscle torque. *Physical Therapy*, 64(9), 1375-1380.
- Hase, T., & Ueo, T. (1999). Acetabular labral tear: arthroscopic diagnosis and treatment. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*, 15(2), 138-141.
- Haversath, M., Hanke, J., Landgraeber, S., Herten, M., Zilkens, C., Krauspe, R., & Jager, M. (2013). The distribution of nociceptive innervation in the painful hip: a histological investigation. *Bone & Joint Journal*, 95 B(6), 770-776. doi:10.1302/0301-620x.95b6.30262
- Hawker, G. A., Mian, S., Kendzerska, T., & French, M. (2011). Measures of adult pain. *Arthritis Care and Research*, 63(SUPPL. 11), S240-S252. doi:10.1002/acr.20543
- Herbert, L. J., Maltais, D. B., Lepage, C., Saulnier, J., Crete, M., & Perron, M. (2011). Isometric muscle strength in youth assessed by hand-held dynamometry: a feasibility, reliability, and validity study. *Pediatric Physical Therapy*, 23(3), 289-299.
- Herbert, R. (2013). *Confidence Interval Calculator*. Retrieved 10 April 2015, 2015, from <http://www.pedro.org.au/english/downloads/confidence-interval-calculator>
- Hicks, G. E., Fritz, J. M., Delitto, A., & McGill, S. M. (2005). Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. *Archives of Physical Medicine and Rehabilitation*, 86(9), 1753-1762.
- Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., . . . Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ (Online)*, 343(7829). doi:10.1136/bmj.d5928
- Holla, J. F. M., van der Leeden, M., Roorda, L. D., Bierma-Zeinstra, S. M. A., Damen, J., Dekker, J., & Steultjens, M. P. M. (2012). Diagnostic accuracy of range of motion measurements in early symptomatic hip and/or knee osteoarthritis. *Arthritis Care and Research*, 64(1), 59-65. doi:10.1002/acr.20645
- Holm, I., Bolstad, B., Lutken, T., Ervik, A., Rokkum, M., & Steen, H. (2000). Reliability of goniometric measurements and visual estimates of hip ROM in patients with osteoarthrosis. *Physiotherapy Research International*, 5(4), 241-248.

- Holmich, P., Holmich, L. R., & Bjerg, A. M. (2004). Clinical examination of athletes with groin pain: an intraobserver and interobserver reliability study. *British Journal of Sports Medicine*, 38(4), 446-451. doi:10.1136/bjism.2003.004754
- Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30(1), 1-15.
- Hosmer, D. W., Lemeshow, S., Sturdivant, R. X., & Ebooks Corporation. (2013). *Applied logistic regression* (EBL1138225 (NhCcYBP)EBL1138225)
- Jacobs, C., Uhl, T. L., Seeley, M., Sterling, W., & Goodrich, L. (2005). Strength and fatigability of the dominant and nondominant hip abductors. *Journal of Athletic Training*, 40(3), 203-206.
- Jaeschke, R., Guyatt, G., & Sackett, D. (1994a). User's guides to the medical literature: III. How to use an article about a diagnostic test: B. What are the results and will they help me in caring for my patients? *JAMA*, 271(9), 703-707.
- Jaeschke, R., Guyatt, G., & Sackett, D. (1994b). Users' guides to the medical literature: III. How to use an article about a diagnostic test A. Are the results of the study valid? *JAMA*, 271(5), 389-391.
- Janssen, J. C., & Le-Ngoc, L. (2009). Intratester reliability and validity of concentric measurements using a new hand-held dynamometer. *Archives of Physical Medicine and Rehabilitation*, 90(9), 1541-1547.
- Jensen, M. P., Turner, J. A., Romano, J. M., & Fisher, L. D. (1999). Comparative reliability and validity of chronic pain intensity measures. *Pain*, 83(2), 157-162. doi:10.1016/s0304-3959(99)00101-3
- Jensen, M. P., Wang, W., Potts, S. L., & Gould, E. M. (2012). Reliability and validity of individual and composite recall pain measures in patients with cancer. *Pain Medicine*, 13(10), 1284-1291.
- Jeon, C. H., Chung, N. S., Lee, Y. S., Son, K. H., & Kim, J. H. (2013). Assessment of hip abductor power in patients with foot drop: a simple and useful test to differentiate lumbar radiculopathy and peroneal neuropathy [Article]. *Spine*, 38(3), 257-263. doi:10.1097/BRS.0b013e318268c8bc
- Joe, G. O., Kovacs, J. A., Miller, K. D., Kelly, G. G., Koziol, D. E., Jones, E. C., . . . Gerber, L. (2002). Diagnosis of avascular necrosis of the hip in asymptomatic HIV-infected patients: clinical correlation of physical examination with magnetic resonance imaging. *Journal of Back and Musculoskeletal Rehabilitation*, 16(4), 135-139.
- Jones, S. R., Carley, S., & Harrison, M. (2003). An introduction to power and sample size estimation. *Emergency Medicine Journal*, 20(5), 453-458.
- Jordan, J. M., Helmick, C. G., Renner, J. B., Luta, G., Dragomir, A. D., Woodard, J., . . . Hochberg, M. C. (2009). Prevalence of hip symptoms and radiographic and symptomatic hip osteoarthritis in African Americans and Caucasians: the Johnston County osteoarthritis project. *Journal of Rheumatology*, 36(4), 809-815. doi:10.3899/jrheum.080677
- Judd, D. L., Thomas, A. C., Dayton, M. R., & Stevens-Lapsley, J. E. (2014). Strength and functional deficits in individuals with hip osteoarthritis compared to healthy, older adults. *Disability and Rehabilitation*, 36(4), 307-312. doi:10.3109/09638288.2013.790491
- Jung, K. A., Restrepo, C., Hellman, M., AbdelSalam, H., Morrison, W., & Parvizi, J. (2011). The prevalence of cam-type femoroacetabular deformity in asymptomatic adults. *Journal of Bone and Joint Surgery (British Volume)*, 93-B(10), 1303-1307. doi:10.1302/0301-620x.93b10.26433

- Jung, S. T., Rowe, S. M., Moon, E. S., Song, E. K., Yoon, T. R., & Seo, H. Y. (2003). Significance of laboratory and radiologic findings for differentiating between septic arthritis and transient synovitis of the hip *Journal of Pediatric Orthopaedics*, *23*(3), 368-372. doi:10.1097/00004694-200305000-00017
- Kampa, R. J., Prasthofer, A., Lawrence-Watt, D. J., & Pattison, R. M. (2007). The internervous safe zone for incision of the capsule of the hip. A cadaveric study. *Journal of Bone and Joint Surgery - Series B*, *89*(7), 971-976. doi:10.1302/0301-620x.89b7.19053
- Kang, A. C. L., Gooding, A. J., Coates, M. H., Goh, T. D., Armour, P., & Rietveld, J. (2010). Computed tomography assessment of hip joints in asymptomatic individuals in relation to femoroacetabular impingement. *American Journal of Sports Medicine*, *38*(6), 1160-1165.
- Karanth, S. S., Springall, D. R., Kuhn, D. M., Levene, M. M., & Polak, J. M. (1991). An immunocytochemical study of cutaneous innervation and the distribution of neuropeptides and protein gene product 9.5 in man and commonly employed laboratory animals [Article]. *American Journal of Anatomy*, *191*(4), 369-383.
- Katoh, M., & Yamasaki, H. (2009). Test-retest reliability of isometric leg muscle strength measurements made using a hand-held dynamometer restrained by a belt: comparisons during and between sessions. *Journal of Physical Therapy Science*, *21*(3), 239-243. doi:10.1589/jpts.21.239
- Keeney, J. A., Peelle, M. W., Jackson, J., Rubin, D., Maloney, W. J., & Clohisy, J. C. (2004). Magnetic resonance arthrography versus arthroscopy in the evaluation of articular hip pathology. *Clinical Orthopaedics and Related Research*(429), 163-169. doi:10.1097/01.bl0.0000150125.34906.7d
- Kellgren, J. H., & Lawrence, J. S. (1957). Radiological assessment of osteo-arthritis. *Annals of the Rheumatic Diseases*, *16*(4), 494-502.
- Kelln, B. M., McKeon, P. O., Gontkof, L. M., & Hertel, J. (2008). Hand-held dynamometry: reliability of lower extremity muscle testing in healthy, physically active, young adults. *Journal of Sport Rehabilitation*, *17*(2), 160-170.
- Kelly, B. T., Weiland, D. E., Schenker, M. L., & Philippon, M. J. (2005). Arthroscopic labral repair in the hip: surgical technique and review of the literature. *Arthroscopy*, *21*(12), 1496-1504. doi:10.1016/j.arthro.2005.08.013
- Kemp, J. L., Makdissi, M., Schache, A. G., Pritchard, M. G., Pollard, T. C. B., & Crossley, K. M. (2014a). Hip chondropathy at arthroscopy: prevalence and relationship to labral pathology, femoroacetabular impingement and patient-reported outcomes. *British Journal of Sports Medicine*, *48*(14), 1102-1107.
- Kemp, J. L., Schache, A. G., Makdissi, M., Pritchard, M. G., Sims, K., & Crossley, K. M. (2014b). Is hip range of motion and strength impaired in people with hip chondrolabral pathology? *Journal of Musculoskeletal Neuronal Interactions*, *14*(3), 334-342. doi:10.1136/bjsports-2013-093312
- Kemp, J. L., Schache, A. G., Makdissi, M., Sims, K. J., & Crossley, K. M. (2013). Greater understanding of normal hip physical function may guide clinicians in providing targeted rehabilitation programmes. *Journal of Science and Medicine in Sport*, *16*(4), 292-296.
- Kim, C., Linsenmeyer, K. D., Vlad, S. C., Guermazi, A., Clancy, M. M., Niu, J., & Felson, D. T. (2014). Prevalence of radiographic and symptomatic hip osteoarthritis in an urban United States community: the Framingham

- osteoarthritis study. *Arthritis and Rheumatology*, 66(11), 3013-3017.  
doi:10.1002/art.38795
- Kim, Y. T., & Azuma, H. (1995). The nerve endings of the acetabular labrum. *Clinical Orthopaedics and Related Research*(320), 176-181.
- Kivlan, B. R., Martin, R. L., & Sekiya, J. K. (2011). Response to diagnostic injection in patients with femoroacetabular impingement, labral tears, chondral lesions, and extra-articular pathology. *Arthroscopy*, 27(5), 619-627.  
doi:10.1016/j.arthro.2010.12.009
- Klassbo, M., Harms-Ringdahl, K., & Larsson, G. (2003). Examination of passive ROM and capsular patterns in the hip. *Physiotherapy Research International*, 8(1), 1-12.
- Klausmeier, V., Lugade, V., Jewett, B. A., Collis, D. K., & Chou, L. S. (2010). Is there faster recovery with an anterior or anterolateral THA? A pilot study. *Clinical Orthopaedics and Related Research*, 468(2), 533-541.  
doi:10.1007/s11999-009-1075-4
- Knapik, J. J., Bauman, C. L., Jones, B. H., Harris, J. M., & Vaughan, L. (1991). Preseason strength and flexibility imbalances associated with athletic injuries in female collegiate athletes. *American Journal of Sports Medicine*, 19(1), 76-81.
- Knols, R. H., Aufdemkampe, G., De Bruin, E. D., Uebelhart, D., & Aaronson, N. K. (2009). Hand-held dynamometry in patients with haematological malignancies: measurement error in the clinical assessment of knee extension strength. *BMC Musculoskeletal Disorders*, 10. doi:10.1186/1471-2474-10-31
- Kocher, M. S., Mandiga, R., Zurakowski, D., Barnewolt, C., & Kasser, J. R. (2004). Validation of a clinical prediction rule for the differentiation between septic arthritis and transient synovitis of the hip in children. *Journal of Bone and Joint Surgery - Series A*, 86(8), 1629-1635.
- Kocher, M. S., Zurakowski, D., & Kasser, J. R. (1999). Differentiating between septic arthritis and transient synovitis of the hip in children: an evidence-based clinical prediction algorithm. *Journal of Bone and Joint Surgery - Series A*, 81(12), 1662-1670.
- Kosek, E., & Ordeberg, G. (2000). Abnormalities of somatosensory perception in patients with painful osteoarthritis normalize following successful treatment. *European Journal of Pain*, 4(3), 229-238. doi:10.1053/eujp.2000.0175
- Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., . . . Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96-106.
- Krause, D. A., Neuger, M. D., Lambert, K. A., Johnson, A. E., DeViny, H. A., & Hollman, J. H. (2014). Effects of examiner strength on reliability of hip-strength testing using a handheld dynamometer. *Journal of Sport Rehabilitation*, 23(1), 56-64. doi:10.1123/jsr.2012-0070
- Krause, D. A., Schlagel, S. J., Stember, B. M., Zoetewey, J. E., & Hollman, J. H. (2007). Influence of lever arm and stabilization on measures of hip abduction and adduction torque obtained by hand-held dynamometry. *Archives of Physical Medicine and Rehabilitation*, 88(1), 37-42.
- Kuhn, J. E., Dunn, W. R., Ma, B., Wright, R. W., Jones, G., Spencer, E. E., . . . Holloway, B. (2007). Interobserver agreement in the classification of rotator cuff tears. *American Journal of Sports Medicine*, 35(3), 437-441.

- Kumar, D., Wyatt, C. R., Lee, S., Nardo, L., Link, T. M., Majumdar, S., & Souza, R. B. (2013). Association of cartilage defects, and other MRI findings with pain and function in individuals with mild-moderate radiographic hip osteoarthritis and controls. *Osteoarthritis and Cartilage*, *21*(11), 1685-1692.
- Kwee, R. M., Kavanagh, E. C., & Adriaensen, M. E. A. P. M. (2013). Normal anatomical variants of the labrum of the hip at magnetic resonance imaging: a systematic review. *European Radiology*, *23*(6), 1694-1710.  
doi:10.1007/s00330-012-2744-3
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.
- Lane, N. E., Nevitt, M. C., Genant, H. K., & Hochberg, M. C. (1993). Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. *Journal of Rheumatology*, *20*(11), 1911-1918.
- Lanyon, P., Muir, K., Doherty, S., & Doherty, M. (2003). Age and sex differences in hip joint space among asymptomatic subjects without structural change: implications for epidemiologic studies. *Arthritis and Rheumatism*, *48*(4), 1041-1046.
- Larkin, B., van Holsbeeck, M., Koueiter, D., & Zaltz, I. (2015). What is the impingement-free range of motion of the asymptomatic hip in young adult males? *Clinical Orthopaedics and Related Research*, *473*(4), 1284-1288.  
doi:10.1007/s11999-014-4072-1
- Larson, C. M., Moreau-Gaudry, A., Kelly, B. T., Thomas Byrd, J. W., Tonetti, J., Lavalley, S., . . . Bedi, A. (2015). Are normal hips being labeled as pathologic? A CT-based method for defining normal acetabular coverage. *Clinical Orthopaedics and Related Research*, *473*(4), 1247-1254.  
doi:10.1007/s11999-014-4055-2
- Laslett, M., Aprill, C. N., McDonald, B., & Young, S. B. (2005). Diagnosis of sacroiliac joint pain: validity of individual provocation tests and composites of tests. *Manual Therapy*, *10*(3), 207-218. doi:10.1016/j.math.2005.01.003
- Laslett, M., & Williams, M. (1994). The reliability of selected pain provocation tests for sacroiliac joint pathology. *Spine*, *19*(11), 1243-1249.
- Laupacis, A., Sekar, N., & Stiell, I. G. (1997). Clinical prediction rules: a review and suggested modifications of methodological standards. *Journal of the American Medical Association*, *277*(6), 488-494.
- Lee, A. J. J., Armour, P., Thind, D., Coates, M. H., & Kang, A. C. L. (2015). The prevalence of acetabular labral tears and associated pathology in a young asymptomatic population. *Bone & Joint Journal*, *97-B*(5), 623-627.  
doi:10.1302/0301-620x.97b5.35166
- Leeftang, M. M. G., Deeks, J. J., Gatsonis, C., & Bossuyt, P. M. M. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine*, *149*(12), 889-897.
- Leibold, M. R., Huijbregts, P., & Jensen, R. (2008). Concurrent criterion-related validity of physical examination tests for hip labral lesions: a systematic review. *Journal of Manual & Manipulative Therapy*, *16*(2), [E24-41].
- Leshner, J. M., Dreyfuss, P., Hager, N., Kaplan, M., & Furman, M. (2008). Hip joint pain referral patterns: a descriptive study. *Pain Medicine*, *9*(1), 22-25.  
doi:10.1111/j.1526-4637.2006.00153.x
- Leunig, M., Beaulé, P. E., & Ganz, R. (2009). The concept of femoroacetabular impingement: current status and future perspectives. *Clinical Orthopaedics and Related Research*, *467*, 616.

- Leunig, M., Beck, M., Stauffer, E., Hertel, R., & Ganz, R. (2000). Free nerve endings in the ligamentum capitis femoris. *Acta Orthopaedica Scandinavica*, 71(5), 452-454.
- Leunig, M., Werlen, S., Ungersböck, A., Ito, K., & Ganz, R. (1997). Evaluation of the acetabular labrum by MR arthrography. *Journal of Bone and Joint Surgery - Series B*, 79(2), 230-234. doi:10.1302/0301-620x.79b2.7288
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine*, 6(7), e1000100.
- Lijmer, J. G., Mol, B. W., Heisterkamp, S., Bossel, G. J., Prins, M. H., Van Der Meulen, J. H. P., & Bossuyt, P. M. M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association*, 282(11), 1061-1066. doi:10.1001/jama.282.11.1061
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255-268.
- Lluch, E., Torres, R., Nijs, J., & Van Oosterwijck, J. (2014). Evidence for central sensitization in patients with osteoarthritis pain: a systematic literature review. *European Journal of Pain (United Kingdom)*, 18(10), 1367-1375. doi:10.1002/j.1532-2149.2014.499.x
- Loureiro, A., Mills, P. M., & Barrett, R. S. (2013). Muscle weakness in hip osteoarthritis: a systematic review. *Arthritis Care and Research*, 65(3), 340-352. doi:10.1002/acr.21806
- Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology*, 63(8), 854-861. doi:10.1016/j.jclinepi.2009.10.002
- Maj, L., Gombar, Y., & Morrison, W. B. (2013). MR imaging of hip infection and inflammation. *Magnetic Resonance Imaging Clinics of North America*, 21(1), 127-139.
- Malliaras, P., Hogan, A., Nawrocki, A., Crossley, K., & Schache, A. (2009). Hip flexibility and strength measures: reliability and association with athletic groin pain. *British Journal of Sports Medicine*, 43(10), 739-744. doi:10.1136/bjism.2008.055749
- Mapp, P. I. (1995). Innervation of the synovium. *Annals of the Rheumatic Diseases*, 54(5), 398-403.
- Mapp, P. I., Kidd, B. L., Gibson, S. J., Terry, J. M., Revell, P. A., Ibrahim, N. B. N., . . . Polak, J. M. (1990). Substance P-, calcitonin gene-related peptide- and C-flanking peptide of neuropeptide Y-immunoreactive fibres are present in normal synovium but depleted in patients with rheumatoid arthritis. *Neuroscience*, 37(1), 143-153. doi:[http://dx.doi.org/10.1016/0306-4522\(90\)90199-E](http://dx.doi.org/10.1016/0306-4522(90)90199-E)
- Martin, H. D., Kelly, B. T., Leunig, M., Philippon, M. J., Clohisy, J. C., Martin, R. L., . . . Safran, M. R. (2010a). The pattern and technique in the clinical evaluation of the adult hip: the common physical examination tests of hip specialists. *Arthroscopy*, 26(2), 161-172.
- Martin, H. D., Shears, S. A., & Palmer, I. J. (2010b). Evaluation of the hip. *Sports Medicine and Arthroscopy Review*, 18(2), 63-75.
- Martin, R. L., Irrgang, J. I., & Sekiya, J. K. (2008). The diagnostic accuracy of a clinical examination in determining intra-articular hip pain for potential hip arthroscopy candidates. *Arthroscopy*, 24(9), 1013-1018.

- Martin, R. L., Kelly, B. T., Leunig, M., Martin, H. D., Mohtadi, N. G., Philippon, M. J., . . . Safran, M. R. (2010c). Reliability of clinical diagnosis in intraarticular hip diseases. *Knee Surgery, Sports Traumatology, Arthroscopy*, *18*(5), 685-690.
- Martin, R. L., & Sekiya, J. K. (2008). The interrater reliability of 4 clinical tests used to assess individuals with musculoskeletal hip pain. *Journal of Orthopaedic and Sports Physical Therapy*, *38*(2), 71-77.
- Maslon, A., Jozwiak, M., Pawlak, M., Modrzewski, T., & Grzegorzewski, A. (2011). Hip joint pain in spastic dislocation: aetiological aspects. *Developmental Medicine and Child Neurology*, *53*(11), 1019-1023.
- Maslowski, E., Sullivan, W., Forster Harwood, J., Gonzalez, P., Kaufman, M., Vidal, A., & Akuthota, V. (2010). The diagnostic validity of hip provocation maneuvers to detect intra-articular hip pathology. *Physical Medicine and Rehabilitation*, *2*(3), 174-181. doi:10.1016/j.pmrj.2010.01.014
- McCarthy, J. C., & Busconi, B. (1995). The role of hip arthroscopy in the diagnosis and treatment of hip disease. *Orthopedics*, *18*(8), 753-756.
- McCarthy, J. C., Noble, P. C., Schuck, M. R., Wright, J., & Lee, J. (2001). The watershed labral lesion. Its relationship to early arthritis of the hip. *Journal of Arthroplasty*, *16*(8 SUPPL.1), 81-87. doi:10.1054/arth.2001.28370
- McDougall, J. J. (2006). Arthritis and pain. Neurogenic origin of joint pain. *Arthritis Research and Therapy*, *8*. doi:10.1186/ar2069
- McGinn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., & Richardson, W. S. (2000). Users' guides to the medical literature XXII: how to use articles about clinical decision rules. *Journal of the American Medical Association*, *284*(1), 79-84.
- McGuire, C. M., MacMahon, P., Byrne, D. P., Kavanagh, E., & Mulhall, K. J. (2012). Diagnostic accuracy of magnetic resonance imaging and magnetic resonance arthrography of the hip is dependent on specialist training of the radiologist. *Skeletal Radiology*, *41*(6), 659-665. doi:10.1007/s00256-011-1266-4
- McLaughlin, K., Eva, K. W., & Norman, G. R. (2014). Reexamining our bias against heuristics. *Advances in Health Sciences Education*, *19*(3), 457-464.
- McNair, P., Prapavassiss, H., Collier, J., Bassett, S., Bryant, A., & Larmer, P. (2007). The Lower-Limb Tasks Questionnaire: an assessment of validity, reliability, responsiveness, and minimal important differences. *Archives of Physical Medicine and Rehabilitation*, *88*(8), 993-1001.
- Michener, L. A., Walsworth, M. K., Doukas, W. C., & Murphy, K. P. (2009). Reliability and diagnostic accuracy of 5 physical examination tests and combination of tests for subacromial impingement. *Archives of Physical Medicine and Rehabilitation*, *90*(11), 1898-1903.
- Mitchell, B., McCrory, P., Brukner, P., O'Donnell, J., Colson, E., & Howells, R. (2003). Hip joint pathology: clinical presentation and correlation between magnetic resonance arthrography, ultrasound, and arthroscopic findings in 25 consecutive cases. *Clinical Journal of Sport Medicine*, *13*(3), 152-156.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, *339*, 332-336. doi:10.1136/bmj.b2535
- Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., . . . de Vet, H. C. W. (2010). Inter-rater agreement and reliability of the COSMIN (Consensus-based Standards for the selection of health status

- Measurement Instruments) checklist. *BMC Medical Research Methodology*, 10(1), 1-11. doi:10.1186/1471-2288-10-82
- Monteiro, S. M., & Norman, G. (2013). Diagnostic reasoning: where we've been, where we're going. *Teaching and Learning in Medicine*, 25(SUPPL.1), S26-S32. doi:10.1080/10401334.2013.842911
- Moraes, M. R. B., Cavalcante, M. L. C., Leite, J. A. D., Macedo, J. N., Sampaio, M. L. B., Jamacaru, V. F., & Santana, M. G. (2011). The characteristics of the mechanoreceptors of the hip with arthrosis. *Journal of Orthopaedic Surgery and Research*, 6(1). doi:10.1186/1749-799x-6-58
- Moreland, J. D., Richardson, J. A., Goldsmith, C. H., & Clase, C. M. (2004). Muscle weakness and falls in older adults: a systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 52(7), 1121-1129. doi:10.1111/j.1532-5415.2004.52310.x
- Morris, S., Dodd, K., & Morris, M. (2008). Reliability of dynamometry to quantify isometric strength following traumatic brain injury. *Brain Injury*, 22(13-14), 1030-1037.
- Mosimann, P. J., Richarme, D., Becce, F., Knoepfli, A.-S., Mino, V., Meuli, R., & Theumann, N. (2012). Usefulness of intra-articular bupivacain and lidocain adjunction in MR or CT arthrography: a prospective study in 148 patients. *European Journal of Radiology*, 81(9), e957-e961. doi:<http://dx.doi.org/10.1016/j.ejrad.2012.06.015>
- Mulligan, E. P., Harwell, J. L., & Robertson, W. J. (2011). Reliability and diagnostic accuracy of the Lachman test performed in a prone position. *Journal of Orthopaedic and Sports Physical Therapy*, 41(10), 749-757. doi:10.2519/jospt.2011.3761
- Myrick, K. M., & Nissen, C. W. (2013). THIRD test: diagnosing hip labral tears with a new physical examination technique. *The Journal for Nurse Practitioners*, 9(8), 501-505. doi:<http://dx.doi.org/10.1016/j.nurpra.2013.06.008>
- Narvani, A. A., Tsiridis, E., Kendall, S., Chaudhuri, R., & Thomas, P. (2003). A preliminary report on prevalence of acetabular labrum tears in sports patients with groin pain. *Knee Surgery, Sports Traumatology, Arthroscopy*, 11(6), 403-408.
- Nathanson, B. H., & Higgins, T. L. (2008). An introduction to statistical methods used in binary outcome modeling. *Seminars in Cardiothoracic and Vascular Anesthesia*, 12(3), 153-166.
- Nepple, J. J., Larson, C. M., Smith, M. V., Kim, Y. J., Zaltz, I., Sierra, R. J., & Clohisy, J. C. (2012). The reliability of arthroscopic classification of acetabular rim labrochondral disease. *American Journal of Sports Medicine*, 40(10), 2224-2229. doi:10.1177/0363546512457157
- Neumann, D. A. (2010). Kinesiology of the hip: a focus on muscular actions. *Journal of Orthopaedic and Sports Physical Therapy*, 40(2), 82-94. doi:10.2519/jospt.2010.3025
- Neumann, G., Mendicuti, A. D., Zou, K. H., Minas, T., Coblyn, J., Winalski, C. S., & Lang, P. (2007). Prevalence of labral tears and cartilage loss in patients with mechanical symptoms of the hip: evaluation using MR arthrography. *Osteoarthritis and Cartilage*, 15(8), 909-917. doi:10.1016/j.joca.2007.02.002
- Niemuth, P. E., Johnson, R. J., Myers, M. J., & Thieman, T. J. (2005). Hip muscle weakness and overuse injuries in recreational runners. *Clinical Journal of Sport Medicine*, 15(1), 14-21. doi:10.1097/00042752-200501000-00004

- Norman, G. R., Coblenz, C. L., Brooks, L. R., & Babcook, C. J. (1992). Expertise in visual diagnosis: a review of the literature. *Academic Medicine*, *67*(10), S78-83.
- Nussbaumer, S., Leunig, M., Glatthorn, J., Stauffacher, S., Gerber, H., & Maffiuletti, N. (2010). Validity and test-retest reliability of manual goniometers for measuring passive hip range of motion in femoroacetabular impingement patients. *BMC Musculoskeletal Disorders*, *11*(1), 194.
- O'Donnell, J., Economopoulos, K., Singh, P., Bates, D., & Pritchard, M. (2014a). The ligamentum teres test: a novel and effective test in diagnosing tears of the ligamentum teres. *American Journal of Sports Medicine*, *42*(1), 138-143. doi:10.1177/0363546513510683
- O'Donnell, J. M., Pritchard, M., Salas, A. P., & Singh, P. J. (2014b). The ligamentum teres—its increasing importance. *Journal of Hip Preservation Surgery*. doi:10.1093/jhps/hnu003
- O'Driscoll, S. L., & Jayson, M. I. (1974). Pain threshold analysis in patients with osteoarthritis of hip. *British Medical Journal*, *3*(5933), 714-715.
- Ochiai, D. H., Adib, F., & Donovan, S. (2011). The twist test: a new test for hip labral pathology. *Arthroscopy*, *27*(5, Supplement), e50. doi:<http://dx.doi.org/10.1016/j.arthro.2011.03.042>
- Ogino, S., Sasho, T., Nakagawa, K., Suzuki, M., Yamaguchi, S., Higashi, M., . . . Moriya, H. (2009). Detection of pain-related molecules in the subchondral bone of osteoarthritic knees [Article]. *Clinical Rheumatology*, *28*(12), 1395-1402. doi:10.1007/s10067-009-1258-0
- Park, M. S., Yoon, S. J., Kim, Y. J., & Chung, W. C. (2014). Hip arthroscopy for femoroacetabular impingement: the changing nature and severity of associated complications over time. *Arthroscopy: The Journal of Arthroscopic and Related Surgery*.
- Pereira, D., Peleteiro, B., Araujo, J., Branco, J., Santos, R. A., & Ramos, E. (2011). The effect of osteoarthritis definition on prevalence and incidence estimates: a systematic review. *Osteoarthritis and Cartilage*, *19*(11), 1270-1285. doi:<http://dx.doi.org/10.1016/j.joca.2011.08.009>
- Petersilge, C. A., Haque, M. A., Petersilge, W. J., Lewin, J. S., Lieberman, J. M., & Buly, R. (1996). Acetabular labral tears: evaluation with MR arthrography. *Radiology*, *200*(1), 231-235.
- Peterson, M. C., Holbrook, J. H., Von Hales, D., Smith, N. L., & Staker, L. V. (1992). Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *Western Journal of Medicine*, *156*(2), 163-165.
- Pewsner, D., Battaglia, M., Minder, C., Marx, A., Bucher, H. C., & Egger, M. (2004). Ruling a diagnosis in or out with "SpPin" and "SnNOut": a note of caution. *British Medical Journal*, *329*(7459), 209-213.
- Philippon, M. J., Briggs, K. K., Yen, Y. M., & Kuppersmith, D. A. (2009). Outcomes following hip arthroscopy for femoroacetabular impingement with associated chondrolabral dysfunction: minimum two-year follow-up. *Journal of Bone and Joint Surgery - Series B*, *91*(1), 16-23. doi:10.1302/0301-620X.91B1.21329
- Philippon, M. J., Maxwell, R., Johnston, T., Schenker, M. L., & Briggs, K. K. (2007). Clinical presentation of femoroacetabular impingement. *Knee Surgery, Sports Traumatology, Arthroscopy*, *15*(8), 1041-1047.

- Phillips, B. A., Lo, S. K., & Mastaglia, F. L. (2000). Muscle force measured using 'break' testing with a hand-held myometer in normal subjects aged 20 to 69 years. *Archives of Physical Medicine and Rehabilitation*, *81*(5), 653-661. doi:10.1053/mr.2000.4413
- Prather, H., Harris-Hayes, M., Hunt, D. M., Steger-May, K., Mathew, V., & Clohisy, J. C. (2010). Reliability and agreement of hip range of motion and provocative physical examination tests in asymptomatic volunteers. *Physical Medicine and Rehabilitation*, *2*(10), 888-895.
- Pua, Y. H., Wrigley, T. W., Cowan, S. M., & Bennell, K. L. (2008). Intrarater test-retest reliability of hip range of motion and hip muscle strength measurements in persons with hip osteoarthritis. *Archives of Physical Medicine and Rehabilitation*, *89*(6), 1146-1154.
- R Core Team. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from <https://http://www.R-project.org>
- Rahman, L. A., Adie, S., Naylor, J. M., Mittal, R., So, S., & Harris, I. A. (2013). A systematic review of the diagnostic performance of orthopedic physical examination tests of the hip. *BMC Musculoskeletal Disorders*, *14*.
- Rasch, A., Byström, A. H., Dalen, N., & Berg, H. E. (2007). Reduced muscle radiological density, cross-sectional area, and strength of major hip and knee muscles in 22 patients with hip osteoarthritis. *Acta Orthopaedica*, *78*(4), 505-510. doi:10.1080/17453670710014158
- Rasch, A., Dalen, N., & Berg, H. E. (2005). Test methods to detect hip and knee muscle weakness and gait disturbance in patients with hip osteoarthritis. *Archives of Physical Medicine and Rehabilitation*, *86*(12), 2371-2376. doi:10.1016/j.apmr.2005.05.019
- Rasch, A., Dalén, N., & Berg, H. E. (2010). Muscle strength, gait, and balance in 20 patients with hip osteoarthritis followed for 2 years after THA. *Acta Orthopaedica*, *81*(2), 183-188. doi:10.3109/17453671003793204
- Ratzlaff, C., Simatovic, J., Wong, H., Li, L., Ezzat, A., Langford, D., . . . Cibere, J. (2013). Reliability of hip examination tests for femoroacetabular impingement. *Arthritis Care and Research*, *65*(10), 1690-1696. doi:10.1002/acr.22036
- Register, B., Pennock, A. T., Ho, C. P., Strickland, C. D., Lawand, A., & Philippon, M. J. (2012). Prevalence of abnormal hip findings in asymptomatic participants: a prospective, blinded study. *American Journal of Sports Medicine*, *40*(12), 2720-2724. doi:10.1177/0363546512462124
- Reijman, M., Hazes, J. M. W., Koes, B. W., Verhagen, A. P., & Bierma-Zeinstra, S. M. A. (2004a). Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: a systematic appraisal. *Annals of the Rheumatic Diseases*, *63*(3), 226-232.
- Reijman, M., Hazes, J. M. W., Pols, H. A. P., Bernsen, R. M. D., Koes, B. W., & Bierma-Zeinstra, S. M. A. (2004b). Validity and reliability of three definitions of hip osteoarthritis: cross sectional and longitudinal approach. *Annals of the Rheumatic Diseases*, *63*(11), 1427-1433.
- Reiman, M. P., Goode, A. P., Cook, C. E., Holmich, P., & Thorborg, K. (2014a). Diagnostic accuracy of clinical tests for the diagnosis of hip femoroacetabular impingement/labral tear: a systematic review with meta-analysis. *British Journal of Sports Medicine*. doi:10.1136/bjsports-2014-094302

- Reiman, M. P., Goode, A. P., Hegedus, E. J., Cook, C. E., & Wright, A. A. (2013). Diagnostic accuracy of clinical tests of the hip: a systematic review with meta-analysis. *British Journal of Sports Medicine*, *47*(14), 893-902.
- Reiman, M. P., Mather, R. C., Hash, T. W., & Cook, C. E. (2014b). Examination of acetabular labral tear: a continued diagnostic challenge. *British Journal of Sports Medicine*, *48*(4), 311-319.
- Reiman, M. P., & Thorborg, K. (2014). Clinical examination and physical assessment of hip joint-related pain in athletes. *The International Journal of Sports Physical Therapy*, *9*(6), 737-755.
- Reiman, M. P., & Thorborg, K. (2015). Femoroacetabular impingement surgery: are we moving too fast and too far beyond the evidence? *British Journal of Sports Medicine*, *49*(12), 782-784. doi:10.1136/bjsports-2014-093821
- Retchford, T. H., Crossley, K. M., Grimaldi, A., Kemp, J. L., & Cowan, S. M. (2013). Can local muscles augment stability in the hip? A narrative literature review. *Journal of Musculoskeletal Neuronal Interactions*, *13*(1), 1-12.
- Reurink, G., Jansen, S. P. L., Bisselink, J. M., Vincken, P. W. J., Weir, A., & Moen, M. H. (2012). Reliability and validity of diagnosing acetabular labral lesions with magnetic resonance arthrography. *Journal of Bone and Joint Surgery*, *94*(18), 1643-1648. doi:10.2106/jbjs.k.01342
- Rho, M., Mautner, K., Nichols, J. T., & Kennedy, D. J. (2013). Image-guided diagnostic injections with anesthetic versus magnetic resonance arthrograms for the diagnosis of suspected hip pain. *Physical Medicine and Rehabilitation*, *5*(9), 795-800. doi:10.1016/j.pmrj.2013.07.008
- Rice, D. A., & McNair, P. J. (2010). Quadriceps arthrogenic muscle inhibition: neural mechanisms and treatment perspectives. *Seminars in Arthritis and Rheumatism*, *40*(3), 250-266. doi:10.1016/j.semarthrit.2009.10.001
- Roaas, A., & Andersson, G. B. J. (1982). Normal range of motion of the hip, knee and ankle joints in male subjects, 30-40 years of age. *Acta Orthopaedica*, *53*(2), 205-208. doi:10.3109/17453678208992202
- Sackett, D. L. (1989). Rules of evidence and clinical recommendations on the use of antithrombotic agents [Article]. *Chest*, *95*(2 SUPPL.), 2S-4S.
- Sackett, D. L. (1992). A primer on the precision and accuracy of the clinical examination. *Journal of the American Medical Association*, *267*(19), 2638-2644. doi:10.1001/jama.267.19.2638
- Salaffi, F., Ciapetti, A., & Carotti, M. (2012). Pain assessment strategies in patients with musculoskeletal conditions. *Reumatismo*, *64*(4), 216-229.
- Salaffi, F., Stancati, A., Silvestri, C. A., Ciapetti, A., & Grassi, W. (2004). Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *European Journal of Pain*, *8*(4), 283-291. doi:10.1016/j.ejpain.2003.09.004
- Sarban, S., Baba, F., Kocabey, Y., Cengiz, M., & Isikan, U. E. (2007). Free nerve endings and morphological features of the ligamentum capitis femoris in developmental dysplasia of the hip. *Journal of Pediatric Orthopaedics Part B*, *16*(5), 351-356. doi:10.1097/01.bpb.0000243830.99681.3e
- Saupe, N., Zanetti, M., Pfirrmann, C. W. A., Wels, T., Schwenke, C., & Hodler, J. (2009). Pain and other side effects after MR arthrography: prospective evaluation in 1085 patients. *Radiology*, *250*(3), 830-838. doi:10.1148/radiol.2503080276
- Saxler, G., Löer, F., Skumavc, M., Pfortner, J., & Hanesch, U. (2007). Localization of SP- and CGRP-immunopositive nerve fibers in the hip joint of patients with

- painful osteoarthritis and of patients with painless failed total hip arthroplasties. *European Journal of Pain*, *11*(1), 67-74.  
doi:10.1016/j.ejpain.2005.12.011
- Schmidt, J., Iverson, J., Brown, S., & Thompson, P. A. (2013). Comparative reliability of the make and break tests for hip abduction assessment. *Physiotherapy Theory and Practice*, *29*(8), 648-657.  
doi:10.3109/09593985.2013.782518
- Schmitt, J. S., & Di Fabio, R. P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology*, *57*(10), 1008-1018.
- Schmitz, M. R., Campbell, S. E., Fajardo, R. S., & Kadrmas, W. R. (2012). Identification of acetabular labral pathological changes in asymptomatic volunteers using optimized, noncontrast 1.5-T magnetic resonance imaging. *American Journal of Sports Medicine*, *40*(6), 1337-1341.  
doi:10.1177/0363546512439991
- Schoth, F., Kraemer, N., Niendorf, T., Hohl, C., Gunther, R. W., & Krombach, G. A. (2008). Comparison of image quality in magnetic resonance imaging of the knee at 1.5 and 3.0 Tesla using 32-channel receiver coils. *European Radiology*, *18*(10), 2258-2264. doi:10.1007/s00330-008-0972-3
- Sherrington, C., & Lord, S. R. (2005). Reliability of simple portable tests of physical performance in older people after hip fracture. *Clinical Rehabilitation*, *19*(5), 496-504.
- Shirai, C., Ohtori, S., Kishida, S., Harada, Y., & Moriya, H. (2009). The pattern of distribution of PGP 9.5 and TNF-alpha immunoreactive sensory nerve fibers in the labrum and synovium of the human hip joint. *Neuroscience Letters*, *450*(1), 18-22. doi:<http://dx.doi.org/10.1016/j.neulet.2008.11.016>
- Siegenthaler, A. M. D., Eichenberger, U. M. D., Schmidlin, K. M. D. D. M. P. H., Arendt-Nielsen, L. P., & Curatolo, M. M. D. (2010). What does local tenderness say about the origin of pain? An investigation of cervical zygapophysial joint pain. *Anesthesia and Analgesia*, *110*(3), 923-927.
- Silvis, M. L., Mosher, T. J., Smetana, B. S., Chinchilli, V. M., Flemming, D. J., Walker, E. A., & Black, K. P. (2011). High prevalence of pelvic and hip magnetic resonance imaging findings in asymptomatic collegiate and professional hockey players. *American Journal of Sports Medicine*, *39*(4), 715-721. doi:10.1177/0363546510388931
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257-268.
- Simel, D. L., Rennie, D., & Bossuyt, P. M. M. (2008). The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. *Journal of General Internal Medicine*, *23*(6), 768-774.  
doi:10.1007/s11606-008-0583-3
- Simel, D. L., Samsa, G. P., & Matchar, D. B. (1991). Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology*, *44*(8), 763-770.
- Sink, E. L., Gralla, J., Ryba, A., & Dayton, M. (2008). Clinical presentation of femoroacetabular impingement in adolescents. *Journal of Pediatric Orthopaedics*, *28*(8), 806-811.

- Smith, A., Jull, G., Schneider, G., Frizzell, B., Hooper, R. A., & Sterling, M. (2014). Cervical radiofrequency neurotomy reduces central hyperexcitability and improves neck movement in individuals with chronic whiplash. *Pain Medicine (United States)*, *15*(1), 128-141. doi:10.1111/pme.12262
- Smith, T. O., Hilton, G., Toms, A. P., Donell, S. T., & Hing, C. B. (2011). The diagnostic accuracy of acetabular labral tears using magnetic resonance imaging and magnetic resonance arthrography: a meta-analysis. *European Radiology*, *21*(4), 863-874. doi:10.1007/s00330-010-1956-7
- Spahn, G., Klinger, H. M., Baums, M., Pinkepank, U., & Hofmann, G. O. (2011). Reliability in arthroscopic grading of cartilage lesions: results of a prospective blinded study for evaluation of inter-observer reliability. *Archives of Orthopaedic and Trauma Surgery*, *131*(3), 377-381. doi:10.1007/s00402-011-1259-8
- Spriet, L. L., Lindinger, M. I., McKelvie, R. S., Heigenhauser, G. J. F., & Jones, N. L. (1989). Muscle glycogenolysis and H<sup>+</sup> concentration during maximal intermittent cycling. *Journal of Applied Physiology*, *66*(1), 8-13.
- Springer, B. A., Gill, N. W., Freedman, B. A., Ross, A. E., Javernick, M. A., & Murphy, K. P. (2009). Acetabular labral tears: diagnostic accuracy of clinical examination by a physical therapist, orthopaedic surgeon, and orthopaedic residents. *North American Journal of Sports Physical Therapy*, *4*(1), 38-45.
- Stanton, T. R., Hancock, M. J., Maher, C. G., & Koes, B. W. (2010). Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions. *Physical Therapy*, *90*(6), 843-853. doi:10.2522/ptj.20090233
- Steultjens, M. P. M., Dekker, J., Van Baar, M. E., Oostendorp, R. A. B., & Bijlsma, J. W. J. (2000). Range of joint motion and disability in patients with osteoarthritis of the knee or hip. *Rheumatology*, *39*(9), 955-961.
- Stochkendahl, M. J., Christensen, H. W., Hartvigsen, J., Vach, W., Haas, M., Hestbaek, L., . . . Bronfort, G. (2006). Manual examination of the spine: a systematic critical literature review of reproducibility. *Journal of Manipulative and Physiological Therapeutics*, *29*(6), 475-475.
- Stockton, K. A., Wrigley, T. V., Mengersen, K. A., Kandiah, D. A., Paratz, J. D., & Bennell, K. L. (2011). Test-retest reliability of hand-held dynamometry and functional tests in systemic lupus erythematosus. *Lupus*, *20*(2), 144-150.
- Stratford, P. W., & Balsor, B. E. (1994). A comparison of make and break tests using a hand-held dynamometer and the Kin-Com. *Journal of Orthopaedic and Sports Physical Therapy*, *19*(1), 28-32.
- Stumbo, T. A., Merriam, S., Nies, K., Smith, A., Spurgeon, D., & Weir, J. P. (2001). The effect of hand-grip stabilization on isokinetic torque at the knee. *Journal of Strength and Conditioning Research*, *15*(3), 372-377. doi:10.1519/1533-4287(2001)015<0372
- Suenaga, E., Noguchi, Y., Jingushi, S., Shuto, T., Nakashima, Y., Miyanishi, K., & Iwamoto, Y. (2002). Relationship between the maximum flexion-internal rotation test and the torn acetabular labrum of a dysplastic hip. *Journal of Orthopaedic Science*, *7*(1), 26-32.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*(1), 13-26.
- Suri, S., Gill, S. E., De Camin, S. M., Wilson, D., McWilliams, D. F., & Walsh, D. A. (2007). Neurovascular invasion at the osteochondral junction and in

- osteophytes in osteoarthritis. *Annals of the Rheumatic Diseases*, 66(11), 1423-1428. doi:10.1136/ard.2006.063354
- Sutlive, T. G., Lopez, H. P., Schnitker, D. E., Yawn, S. E., Halle, R. J., Mansfield, L. T., . . . Childs, J. D. (2008). Development of a clinical prediction rule for diagnosing hip osteoarthritis in individuals with unilateral hip pain. *Journal of Orthopaedic and Sports Physical Therapy*, 38(9), 542-550.
- Sutter, R., Zubler, V., Hoffmann, A., Mamisch-Saupe, N., Dora, C., Kalberer, F., . . . Pfirrmann, C. W. A. (2014). Hip MRI: How useful is intraarticular contrast material for evaluating surgically proven lesions of the labrum and articular cartilage? *American Journal of Roentgenology*, 202(1), 160-169. doi:10.2214/AJR.12.10266
- Szadek, K. M., Hoogland, P. V. J. M., Zuurmond, W. W. A., De Lange, J. J., & Perez, R. S. G. M. (2010). Possible nociceptive structures in the sacroiliac joint cartilage: an immunohistochemical study. *Clinical Anatomy*, 23(2), 192-198. doi:10.1002/ca.20908
- Takeshita, M., Nakamura, J., Ohtori, S., Inoue, G., Orita, S., Miyagi, M., . . . Takahashi, K. (2012). Sensory innervation and inflammatory cytokines in hypertrophic synovia associated with pain transmission in osteoarthritis of the hip: a case-control study [Article]. *Rheumatology*, 51(10), 1790-1795.
- Tang, H. (2007). Diagnostic greed: using pictures to highlight diagnostic errors. *Postgraduate Medical Journal*, 83(977), 209-210. doi:10.1136/pgmj.2006.053280
- Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research*, 21(4), 651-657. doi:10.1007/s11136-011-9960-1
- Thakur, M., Dickenson, A. H., & Baron, R. (2014). Osteoarthritis pain: nociceptive or neuropathic? [Perspectives]. *Nature Reviews Rheumatology*, 10(6), 374-380. doi:10.1038/nrrheum.2014.47
- Thorborg, K., Bandholm, T., Schick, M., Jensen, J., & Holmich, P. (2013). Hip strength assessment using handheld dynamometry is subject to intertester bias when testers are of different sex and strength. *Scandinavian Journal of Medicine and Science in Sports*, 23(4), 487-493. doi:10.1111/j.1600-0838.2011.01405.x
- Thorborg, K., Coupe, C., Petersen, J., Magnusson, S. P., & Holmich, P. (2011a). Eccentric hip adduction and abduction strength in elite soccer players and matched controls: a cross-sectional study. *British Journal of Sports Medicine*, 45(1), 10-13. doi:10.1136/bjism.2009.061762
- Thorborg, K., Petersen, J., Magnusson, S. P., & Hölmich, P. (2010). Clinical assessment of hip strength using a hand-held dynamometer is reliable. *Scandinavian Journal of Medicine and Science in Sports*, 20(3), 493-501.
- Thorborg, K., Serner, A., Petersen, J., Moller Madsen, T., Magnusson, P., & Holmich, P. (2011b). Hip adduction and abduction strength profiles in elite soccer players: implications for clinical evaluation of hip adductor muscle recovery after injury. *American Journal of Sports Medicine*, 39(1), 121-126. doi:10.1177/0363546510378081
- Tibor, L. M., & Sekiya, J. K. (2008). Differential diagnosis of pain around the hip joint. *Arthroscopy*, 24(12), 1407-1421.

- Tijssen, M., van Cingel, R., Willemsen, L., & de Visser, E. (2012). Diagnostics of femoroacetabular impingement and labral pathology of the hip: a systematic review of the accuracy and validity of physical tests. *Arthroscopy*, *28*(6), 860-871.
- Treede, R. D., Jensen, T. S., Campbell, J. N., Cruccu, G., Dostrovsky, J. O., Griffin, J. W., . . . Serra, J. (2008). Neuropathic pain: redefinition and a grading system for clinical and research purposes. *Neurology*, *70*(18), 1630-1635.  
doi:10.1212/01.wnl.0000282763.29778.59
- Troelsen, A., Mechlenburg, I., Gelineck, J., Bolvig, L., Jacobsen, S., & Soballe, K. (2009). What is the role of clinical tests and ultrasound in acetabular labral tear diagnostics? *Acta Orthopaedica*, *80*(3), 314-318.
- Tseng, Y. L., Wang, W. T. J., Chen, W. Y., Hou, T. J., Chen, T. C., & Lieu, F. K. (2006). Predictors for the immediate responders to cervical manipulation in patients with neck pain. *Manual Therapy*, *11*(4), 306-315.  
doi:10.1016/j.math.2005.08.009
- Tyler, T. F., Nicholas, S. J., Campbell, R. J., & McHugh, M. P. (2001). The association of hip strength and flexibility with the incidence of adductor muscle strains in professional ice hockey players. *American Journal of Sports Medicine*, *29*(2), 124-128.
- van Dijk, C. N., Reilingh, M. L., Zengerink, M., & van Bergen, C. J. A. (2010). Osteochondral defects in the ankle: why painful? *Knee Surgery, Sports Traumatology, Arthroscopy*, *18*(5), 570-580. doi:10.1007/s00167-010-1064-x
- Verrall, G. M., Slavotinek, J. P., Barnes, P. G., & Fon, G. T. (2005). Description of pain provocation tests used for the diagnosis of sports-related chronic groin pain: relationship of tests to defined clinical (pain and tenderness) and MRI (pubic bone marrow oedema) criteria. *Scandinavian Journal of Medicine and Science in Sports*, *15*(1), 36-42.
- Versi, E. (1992). 'Gold standard' is an appropriate term *British Medical Journal*, *305*(6846), 187.
- Vicenzino, B., Collins, N., Cleland, J., & McPoil, T. (2010). A clinical prediction rule for identifying patients with patellofemoral pain who are likely to benefit from foot orthoses: a preliminary determination. *British Journal of Sports Medicine*, *44*(12), 862-866.
- Vicenzino, B., Smith, D., Cleland, J., & Bisset, L. (2009). Development of a clinical prediction rule to identify initial responders to mobilisation with movement and exercise for lateral epicondylalgia. *Manual Therapy*, *14*(5), 550-554.
- Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, *165*(6), 710-718. doi:10.1093/aje/kwk052
- Wainner, R. S., Fritz, J. M., Irrgang, J. J., Boninger, M. L., Delitto, A., & Allison, S. (2003). Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. *Spine*, *28*(1), 52-62.  
doi:10.1097/00007632-200301010-00014
- Wang, C. Y., Olson, S. L., & Protas, E. J. (2002). Test-retest strength reliability: hand-held dynamometry in community-dwelling elderly fallers. *Archives of Physical Medicine and Rehabilitation*, *83*(6), 811-815.  
doi:10.1053/apmr.2002.32743
- Wang, W. G., Yue, D. B., Zhang, N. F., Hong, W., & Li, Z. R. (2011). Clinical diagnosis and arthroscopic treatment of acetabular labral tears. *Orthopaedic Surgery*, *3*(1), 28-34.

- Ward, S. R., Winters, T. M., & Blemker, S. S. (2010). The architectural design of the gluteal muscle group: implications for movement and rehabilitation. *Journal of Orthopaedic and Sports Physical Therapy*, *40*(2), 95-102. doi:10.2519/jospt.2010.3302
- Weingarten, T. N., Watson, J. C., Hooten, W. M., Wollan, P. C., Melton Iii, L. J., Locketz, A. J., . . . Yawn, B. P. (2007). Validation of the S-LANSS in the community setting. *Pain*, *132*(1-2), 189-194.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231-240. doi:10.1519/15184.1
- Weir, J. P., Wagner, L. L., & Housh, T. J. (1994). The effect of rest interval length on repeated maximal bench presses. *Journal of Strength and Conditioning Research*, *8*(1), 58-60.
- Wells, G. A., Shea, B. J., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2015). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Retrieved 24 July 2015, 2015, from [http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.asp](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp)
- White, S. G., McNair, P., Laslett, M., & Hing, W. (2015). Do patients undergoing physical testing report pain intensity reliably? *Arthritis Care and Research*, *67*(6), 873-879. doi:10.1002/acr.22530
- Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy in systematic reviews. *BMC Medical Research Methodology*, *3*.
- Whiting, P., Rutjes, A., Westwood, M., Mallett, S., Deeks, J., Reitsma, J., . . . Bossuyt, P. (2011). Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529-536.
- Whiting, P., Weswood, M., Rutjes, A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2006). Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology*, *6*(1), 9.
- Wieners, G., Detert, J., Streitparth, F., Pech, M., Fischbach, F., Burmester, G., . . . Bruhn, H. (2007). High-resolution MRI of the wrist and finger joints in patients with rheumatoid arthritis: comparison of 1.5 Tesla and 3.0 Tesla. *European Radiology*, *17*(8), 2176-2182. doi:10.1007/s00330-006-0539-0
- Woolf, A. D. (2003). History and physical examination. *Best Practice & Research: Clinical Rheumatology*, *17*(3), 381-402.
- Wright, A. A., Cook, C. E., Flynn, T. W., Baxter, G. D., & Abbott, J. H. (2011). Predictors of response to physical therapy intervention in patients with primary hip osteoarthritis. *Physical Therapy*, *91*(4), 510-524.
- Wyrwich, K. W. (2004). Minimal important difference thresholds and the standard error of measurement: is there a connection? *Journal of Biopharmaceutical Statistics*, *14*(1), 97-110. doi:10.1081/bip-120028508
- Wyss, T. F., Clark, J. M., Weishaupt, D., & Nötzli, H. P. (2007). Correlation between internal rotation and bony anatomy in the hip. *Clinical Orthopaedics and Related Research*(460), 152-158. doi:10.1097/BLO.0b013e3180399430
- Yeung, S. S., Suen, A. M. Y., & Yeung, E. W. (2009). A prospective cohort study of hamstring injuries in competitive sprinters: preseason muscle imbalance as a possible risk factor. *British Journal of Sports Medicine*, *43*(8), 589-594. doi:10.1136/bjism.2008.056283

- Youdas, J. W., Madson, T. J., & Hollman, J. H. (2010). Usefulness of the Trendelenburg test for identification of patients with hip joint osteoarthritis. *Physiotherapy Theory and Practice, 26*(3), 184-194. doi:10.3109/09593980902750857
- Youdas, J. W., Mraz, S. T., Norstad, B. J., Schinke, J. J., & Hollman, J. H. (2008). Determining meaningful changes in hip abductor muscle strength obtained by handheld dynamometry. *Physiotherapy Theory and Practice, 24*(3), 215-220.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35.
- Young, S., & Aprill, C. (2000). Characteristics of a mechanical assessment for chronic lumbar facet pain. *The Journal of Manual & Manipulative Therapy, 8*(2), 78-84.
- Young, S., Aprill, C., & Laslett, M. (2003). Correlation of clinical examination characteristics with three sources of chronic low back pain. *The Spine Journal, 3*, 460-465.
- Zifchock, R. A., Davis, I., Higginson, J., McCaw, S., & Royer, T. (2008). Side-to-side differences in overuse running injury susceptibility: a retrospective study. *Human Movement Science, 27*(6), 888-902. doi:10.1016/j.humov.2008.03.007

## Appendices

## Appendix 1. Ethical approval for interview and reliability studies



**MEMORANDUM**  
Auckland University of Technology Ethics Committee (AUTEC)

---

To: Peter McNair  
 From: **Madeline Banda** Executive Secretary, AUTEC  
 Date: 14 May 2010  
 Subject: Ethics Application Number 10/44 **The reliability of clinical tests used in the assessment of hip joint pain.**

---

Dear Peter

Thank you for providing written evidence as requested. I am pleased to advise that it satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC) at their meeting on 12 April 2010 and that on 7 May 2010 I approved your ethics application. This delegated approval is made in accordance with section 5.3.2.3 of AUTEC's *Applying for Ethics Approval: Guidelines and Procedures* and is subject to endorsement at AUTEC's meeting on 14 June 2010.

Your ethics application is approved for a period of three years until 7 May 2013.

I advise that as part of the ethics approval process, you are required to submit the following to AUTEC:

- A brief annual progress report using form EA2, which is available online through <http://www.aut.ac.nz/research/research-ethics>. When necessary this form may also be used to request an extension of the approval at least one month prior to its expiry on 7 May 2013;
- A brief report on the status of the project using form EA3, which is available online through <http://www.aut.ac.nz/research/research-ethics>. This report is to be submitted either when the approval expires on 7 May 2013 or on completion of the project, whichever comes sooner;

It is a condition of approval that AUTEC is notified of any adverse events or if the research does not commence. AUTEC approval needs to be sought for any alteration to the research, including any alteration of or addition to any documents that are provided to participants. You are reminded that, as applicant, you are responsible for ensuring that research undertaken under this approval occurs within the parameters outlined in the approved application.

Please note that AUTEC grants ethical approval only. If you require management approval from an institution or organisation for your research, then you will need to make the arrangements necessary to obtain this. Also, if your research is undertaken within a jurisdiction outside New Zealand, you will need to make the arrangements necessary to meet the legal and ethical requirements that apply within that jurisdiction.

When communicating with us about this application, we ask that you use the application number and study title to enable us to provide you with prompt service. Should you have any further enquiries regarding this matter, you are welcome to contact Charles Grinter, Ethics Coordinator, by email at [ethics@aut.ac.nz](mailto:ethics@aut.ac.nz) or by telephone on 921 9999 at extension 8860.

On behalf of the AUTEC and myself, I wish you success with your research and look forward to reading about it in your reports.

Yours sincerely



**Madeline Banda**  
 Executive Secretary  
 Auckland University of Technology Ethics Committee

Cc: Steven White [steve.white@aut.ac.nz](mailto:steve.white@aut.ac.nz)

---

From the desk of  
 Madeline Banda  
 Executive Secretary  
 AUTEC

Private Bag 92006, Auckland 1142  
 New Zealand  
 E-mail: [ethics@aut.ac.nz](mailto:ethics@aut.ac.nz)

Tel: 64 9 921 9999  
 ext 8044  
 Fax: 64 9 921 9925  
 page 1 of 1

## Appendix 2. Reliability study information sheet

# Participant Information Sheet



Date Information Sheet Produced: 22 March 2010

### Project Title

The reliability of clinical tests used in the assessment of hip joint pain

### An Invitation

I am Steven White, a senior lecturer and PhD candidate at AUT University and a qualified, registered physiotherapist. You are invited to participate in a research study. This study will contribute to my PhD.

Participation is completely voluntary (your choice) and you may withdraw from the study at any time prior to the completion of data collection without giving a reason or being disadvantaged. Your participation in this study will be stopped should any harmful effects appear.

### What is the purpose of this research?

This study aims to investigate the reliability of tests commonly used in a clinical setting to diagnose hip joint problems. This will help physiotherapists and doctors to improve their ability to determine the cause of pain felt in the hip region without the need for more expensive and invasive investigations like x-rays and injections into the joint.

If you agree, data collected from this study will be used for further studies planned to follow on from the reliability study. Specifically a study that determines how accurate hip joint tests are in the identification of pain that originates from within the hip and then another which will evaluate the effectiveness of physiotherapy treatment of hip joint pain. No material that could personally identify you will be used in any reports on this study unless your personal approval is given for the dissemination of results to specific persons (please see the section below titled "How will my privacy be protected?" for more information on privacy issues).

### Are you eligible to participate in this project?

If you are between the age of 20 and 80 years old and have pain mainly in the groin or deep buttock region on one side, which has been present for a minimum of one month you are eligible to participate in this study.

You are not eligible to participate in this study if you:

- are unable to speak English
- have a history of hip joint surgery
- have consulted a medical or other health professional for low back pain within the 12 months prior to your current episode of 'hip' pain
- have evidence of nerve compression in your back or legs (e.g. pins & needles, muscle weakness)
- have any systemic disease or illness
- are pregnant
- have severe pain

### What will happen in this research?

This study involves two assessment sessions, the first takes about 90 minutes, the second about 1 hour. All measurements will be undertaken at the North Shore Campus, AUT University. A step by step description of what will happen is described below.

#### A: Subject Preparation

You will be met by the researcher and the experimental procedure will be explained. Your informed consent will be gained. You will be required to complete standardised forms that give background detail about you such as age, occupation and levels of activity as well as detail specific to your hip pain. This will include

information about the exact site of your pain, the intensity of your pain and how it effects your day to day activities. You will be given an opportunity in private to change into your shorts and t-shirt (or similar) so that the researcher can test your hip in a manner that you feel safe and comfortable. If you wish to have a chaperone present during the testing you are welcome to bring someone along with you or to ask the researcher to arrange for an appropriate person for you.

#### B: Equipment

The researcher will use equipment that measures the range of movement of your hip and the amount of force you can generate during tests. This equipment is not invasive and will not cause any damage to you or your clothes.

#### C: Test procedures

The researcher will perform a number of tests on your hip that aim to determine the joint range of motion and hip muscle strength. Additionally, some tests are performed to see if they reproduce any discomfort or pain that you consider to be very similar to that which you normally get with your hip. These tests are all tests that are commonly used by physiotherapists and doctors for patients with hip joint pain. It is very likely that you have had most of them performed on you already if you have seen a medical professional because of your hip pain.

Some tests will be performed three times so that the researcher can later average the results over three trials. The researcher will measure and record detail about range of movement, strength and reproduction of any discomfort or pain that you report. You will be asked to 'score' the level of any discomfort/pain that you feel, on a scale that has been tested and shown to be a valid measurement scale for pain. This sequence of tests will be performed a second time after a 5-10 mins break so that the researcher can compare the results from both testing sessions to see how consistent the tests actually are. The same tests will be repeated **approximately 5 days later** when you return for your second session. This is important as it will help the researcher to see how consistent the results are over a number of days instead of just on one day.

#### **What are the discomforts and risks?**

There is a risk of some soreness remaining after the completion of the tests. This is no different to how you might feel after going to your doctor or physiotherapist for an examination to get a diagnosis of your hip pain so that he/she can advise in regards to the best treatment for you. It is not expected that you will have any significant soreness. Any soreness you may have is likely to settle on its own within 1-2 days.

#### **How will these discomforts and risks be alleviated?**

The researcher has 30 years of experience as a physiotherapist specialising in treating musculoskeletal pain and is well able to recognise any signs that would suggest that you might have a hip that is likely to be aggravated by the test procedures. This is part of the reason why you are asked to complete the initial questionnaires mentioned above. There are a number of details that help to identify those who might experience this type of flare up in symptoms including:

- What, if any, reaction you have had to previous testing
- How you react to day to day activities and/or sporting/occupational activities
- How long any flare-ups that you have experienced in the past have taken to settle.

**Importantly, during the test session, should you feel any excessive discomfort, please advise the researcher and he will stop the performance of that test. If you wish, you can be withdrawn from the rest of the experiment. If you feel any discomfort following the study, you should inform the researcher who will provide you with advice and options regarding the management of any discomfort.**

#### **What are the benefits?**

The results of this study may help to improve the diagnosis and management of people with hip joint pain. It may well lead to earlier diagnosis of hip problems and in doing so decrease the number of complications that result from delayed diagnosis. It may also decrease the need for more expensive and invasive testing procedures.

**What compensation is available for injury or negligence?**

In the unlikely event of a physical injury as a result of your participation in this study, rehabilitation and compensation for injury by accident may be available from the Accident Compensation Corporation, providing the incident details satisfy the requirements of the law and the Corporation's regulations.

**How will my privacy be protected?**

Any information we collect will not be able to be identified as belonging to you. All data collected will be coded so that you will only be identified by a number. The researchers will be the only people who have access to this information. All information will be kept in a secure room and in a locked filing cabinet for a 10 year period and then be destroyed.

**What are the costs of participating in this research?**

There is no financial cost to you to participate in this research. It will take approximately 3 hours of your time.

**What opportunity do I have to consider this invitation?**

You have two weeks to decide whether you wish to take part in the study. You have a right to choose not to participate. If you agree to take part you are free to withdraw from the study at any time prior to the completion of data collection, without having to give a reason.

**How was I chosen to participate in this research?**

You are a person with hip joint pain and you will have responded to one of the advertisements the researcher placed in waiting rooms of selected medical professionals or in a local paper. The adverts in the medical rooms were placed there because it was expected that people with hip pain might be more likely to see advertising and information about the study in that environment than elsewhere.

**How do I agree to participate in this research?**

If you agree to participate in the study please complete the attached consent form.

**Will I receive feedback on the results of this research?**

If you wish to have a copy of the results of this research, please indicate this on the consent form. Results will be available after the study is completed and published.

**What do I do if I have concerns about this research?**

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor, Peter McNair, [peter.mcnair@aut.ac.nz](mailto:peter.mcnair@aut.ac.nz), and a 921 9999 ext 7143

Concerns regarding the conduct of the research should be notified to the Executive Secretary, AUTEK, Madeline Banda, [madeline.banda@aut.ac.nz](mailto:madeline.banda@aut.ac.nz), 921 9999 ext 8044.

**Whom do I contact for further information about this research?**

Please feel free to contact the researcher if you have any questions about this study.

**Researcher Contact Details:**

Steve White; [steve.white@aut.ac.nz](mailto:steve.white@aut.ac.nz) ph 09 921 9999 ext 7073.

**Approved by the Auckland University of Technology Ethics Committee on 14 May 2010  
AUTEK Reference number 10/44**

*Note: The Participant should retain a copy of this form.*

### Appendix 3. Consent form for reliability study

<h1>Consent Form</h1>	 <p><b>AUT</b> UNIVERSITY TE WĀNANGA ARONUI O TAMAKI MAKAU RAU</p>
-----------------------	---

*Project title:*                    **The reliability of clinical tests used in the assessment of hip joint pain**

*Project Supervisor:*        **Professor Peter McNair**

*Researcher:*                    **Mr Steven White**

*Please Circle*

Have you read and understood the information provided about this research project in the information sheet dated 22 March 2010?	Yes	No
Have you had an opportunity to ask questions and have them answered?	Yes	No
Do you understand that you may withdraw yourself or any information that you have provided from this project at any time prior to completion of data collection, without being disadvantaged in any way?	Yes	No
Do you have pain mainly in the groin or deep buttock region on one side, which has been present for a minimum of one month?	Yes	No
Are you 20 years old or over?	Yes	No
Do you have any of the following:		
o a history of hip joint surgery?	Yes	No
o any severe pain?	Yes	No
o any systemic disease or illness ?	Yes	No
o any musculoskeletal injury that may affect your performance other than your painful hip?	Yes	No
o any pins & needles or weakness in your legs?	Yes	No
Have you consulted a medical or other health professional for low back pain within the last 12 months?	Yes	No
Are you pregnant?	Yes	No
<b>Do you consent to taking part in this research?</b>	Yes	No
Do you wish to receive a summary of the research results?	Yes	No
Do you consent to the use of your data by this researcher for the follow up studies outlined in the information sheet dated 22 March 2010, providing they are given ethical approval by an appropriate Health & Disability Ethics committee?	Yes	No
Do you give permission to be contacted regarding the possibility of participating in further research studies into the treatment of hip pain at the conclusion of this study?	Yes	No
Full Name		
Address		
Phone	H	W
		Mob

Participant signature: .....

Date:

**Approved by the Auckland University of Technology Ethics Committee on 14 May 2010**  
**AUTEC Reference number 10/44**

*Researcher Copy*

## Appendix 4. Screening and baseline questionnaire for reliability study

Participant ID \_\_\_\_\_



# Medical Screening Questionnaire

### General Screening Questions:

	Please Tick	
	YES	NO
Have you recently had a fever?		
Have you recently taken antibiotics or other medicines for an infection?		
Have you recently had surgery?		
Have you recently had an injection into any joint?		
Have you recently had a cut or scrape or open wound?		
Have you been diagnosed with an immune-suppressive disorder?		
Have you used IV drugs?		
Do you have a history of cancer?		
Have you recently lost weight for no apparent reason?		
Have you recently felt unwell or unusually tired?		
Have you been waking with pain at night that has not let you get back to sleep reasonably quickly?		
Have you recently taken a long car, bus, plane or train ride?		
Have you recently been bedridden for any reason?		
Have you ever had surgery to your painful hip or back?		
Do you have any numbness or tingling anywhere in your legs?		
Have you ever taken any steroids?		

Participant ID \_\_\_\_\_

**Medications:**

Please list ALL medications you are currently taking:

---



---



---

**General Health:**Do you have any other medical conditions for which you receive treatment? Yes  No 

If yes, please describe that condition(s):

---



---

**Smoking History:**

Cigarettes per day:

Non-smoker  1-10  11-20  21-30  30+ **Surgical History:**

Please list any surgery for which you had a general anaesthetic:

Date:      Surgery:

\_\_\_/\_\_\_/\_\_\_ \_\_\_\_\_

\_\_\_/\_\_\_/\_\_\_ \_\_\_\_\_

\_\_\_/\_\_\_/\_\_\_ \_\_\_\_\_

**Allergies:**

Please list all known allergies:

---



---



---

Participant ID \_\_\_\_\_



## BASELINE SCREENING & DATA COLLECTION

### Background Details:

1. Name \_\_\_\_\_

Office  
Use

2. Address \_\_\_\_\_

3. Contact Phone numbers

a. Home \_\_\_\_\_ Work \_\_\_\_\_ Mobile \_\_\_\_\_

M F  
0 14. Gender Male  Female  Date of Birth \_\_\_\_\_

### Previous Assessment Details:

Y N  
1 05. Have you had your hip examined by a medical professional? Yes  No 

If No, go directly to question 10

Y N  
1 06. If yes, did this examination hurt your hip significantly at the time? Yes  No 

If No, go directly to question 10

0 1 2 3 4  
na=100

7. If yes, approximately how long were you sore for after the examination?

A few minutes  Several Hours  Overnight  1-2 days  More than 3 days Y N na  
1 0 1008. Did you need to take any medication for the soreness you referred to in Q6? Yes  No Y N na  
1 0 100

9. If yes, please provide detail of what medication you took and for how long.

### Irritability:

Y N S  
1 0 210. Is it easy for you to aggravate your hip pain? Yes  No  Sometimes 0 1 2 3 4  
na=100

11. If you aggravate your hip pain, how long does it normally take to settle again?

A few minutes  Several Hours  Overnight  1-2 days  More than 3 days Y N S  
1 0 212. Do you take any medication for your hip pain? Yes  No  Sometimes 1=Analg  
2=NSAIDS  
3=Other

13. If yes (or sometimes), please provide detail of what you normally take and how often.

Y N  
1 014. Are you sometimes totally free of pain, even if only for a few minutes? Yes  No

Participant ID \_\_\_\_\_

**Pain Severity:**

When answering these questions, think only of the pain you are experiencing in relation to your hip.

Please circle one number for each of the following 4 questions:

On average, how bad has your pain been:

- |  | No<br>Pain | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Pain as Bad<br>as it Can Be |
|--|------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|
| 15. In the morning over the past <u>two days</u> ?   |            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |
| 16. In the afternoon over the past <u>two days</u> ? |            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |
| 17. In the evening over the past <u>two days</u> ?   |            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |
| 18. With activity over the past <u>two days</u> ?    |            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |

Office  
Use  
Y N  
1 0

19. Would you say that the severity of pain you have had over the last two days (as you have indicated in questions 15 to 18 above) is similar to what you 'typically' have over a two day period?

Yes  No

20. If you answered No to the question above (Q19), please circle the number below that best describes your average level of your pain over a 'typical' two days.

Na=100  
Score=

- | No<br>Pain | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Pain as Bad<br>as it Can Be |
|------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|
|            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |

Please circle one number for each of the following 2 questions:

21. What is the WORST/HIGHEST level of pain you have experienced at ANY time over the last month

Score=

- | No<br>Pain | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Pain as Bad<br>as it Can Be |
|------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|
|            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |

22. What is the LOWEST level of pain you have experienced at ANY time over the last month

Score=

- | No<br>Pain | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Pain as Bad<br>as it Can Be |
|------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|
|            | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |                             |

**Global Disability Rating**

23. How much has your hip problem affected your normal daily activity in the last week?  
(Please Tick the appropriate box)

- |                             |                             |                             |                             |                             |                             |                             |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| 1. <input type="checkbox"/> | 2. <input type="checkbox"/> | 3. <input type="checkbox"/> | 4. <input type="checkbox"/> | 5. <input type="checkbox"/> | 6. <input type="checkbox"/> | 7. <input type="checkbox"/> |
| No<br>Disability            | Almost no<br>Disability     | Minimal<br>Disability       | Some<br>Disability          | Moderate<br>Disability      | A lot of<br>Disability      | Maximum<br>Disability       |

Participant ID \_\_\_\_\_

Office Use  
1 2 3 4

**Type of Symptoms:**

24. What is the MAIN/WORST problem with your hip (tick only ONE):

Pain       Stiffness       Weakness       Leg giving way

Other (please describe): \_\_\_\_\_

S A O  
1 2 3

25. How would you describe your MAIN pain?

Sharp pain       Aching (deep) pain       Other (please describe): \_\_\_\_\_

Y N  
1 0

26. Do you experience any of the following symptoms?

Crepitus (grinding/creaking or similar).      Yes       No

Clicking or clunking that is painful.      Yes       No

Clicking or clunking that is not associated with pain.      Yes       No

Locking of the hip.      Yes       No

Tingling/numbness anywhere in your legs.      Yes       No

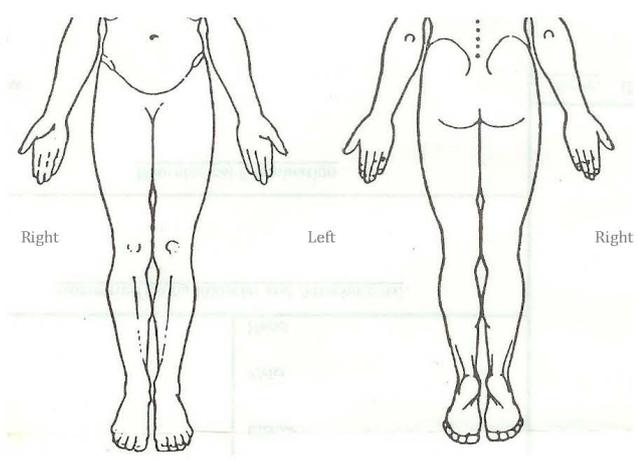
Giving way of your leg because of pain or weakness.      Yes       No

Other (describe): \_\_\_\_\_

**Location of Symptoms:**

27. Please shade in on the diagram below the areas where you feel your pain or other symptoms. Use the letters in the 'Key' to describe the type of pain you have.

28. Please indicate with a cross (X) the location of your **main/worst** hip pain/symptoms.



KEY	
S	= sharp or stabbing pain
A	= aching pain
B	= burning or coldness
PN	= pins & needles or numbness

1  
2  
3  
4

Office Use	Y N 1 0
1= Groin	
2= Upper thigh	
3= Ant thigh	
4= Knee	
5= Mid-buttock	
6= Low buttock	
7= Upper hamstrings	
8= Post thigh	
9= Trochanteric	
10= Other	

Participant ID \_\_\_\_\_

Office Use  
# Mths

### Onset of Symptoms

29. Your **current** hip symptoms have been present since: \_\_\_\_\_ (dd/mm/yy)  
(If you cannot remember the exact date, please give the closest **estimate** for the month/year)

T S O U  
1 2 3 4

30. Describe how your **current** hip symptoms started:  
Trauma       Strain/Injury       Overuse       Unknown   
e.g. fall, impact, accident      e.g. strained lifting      e.g. repetitive action

Briefly describe what you feel caused your current hip pain:  
-----

1 2 3

31. If your current pain resulted from a specific event, how long after that event did you **FIRST** notice the pain?

Immediately       Within 48 hours       After more than 48 hours

If the pain was not immediate, describe what you were doing when you **FIRST** noticed the pain:  
-----

1 2 3  
4 5 6 7

32. From the time your symptoms **FIRST** started, when did you **FIRST** consult a medical professional (Dr or physio etc) about the problem?

Within 1 week       Within 1 month       Within 3 months   
Within 6 months       Within 12 months       More than 12 months       Never

Y N  
1 0

### Previous Hip Pain

# Mths

33. Have you had **previous** problems with this (same) hip?      Yes       No       If 'no' go to **Q 39**

34. If yes, when did you **FIRST** have problems with this hip? \_\_\_\_/\_\_\_\_/\_\_\_\_ (approximate date)

T S O U  
1 2 3 4  
na=100

35. How did the previous problem **first** start?  
Trauma       Strain/Injury       Overuse       Unknown   
e.g. fall, impact, accident      e.g. strained lifting      e.g. repetitive action

Briefly describe what you feel caused your previous hip pain:  
-----

Y N na  
1 0 100

36. Did the problem **fully** resolve before the current episode?      Yes       No

Y N na  
1 0 100

37. Did you consult a medical professional about the previous hip problem?      Yes       No

X U M O  
1 2 3 4  
na=100

38. Did you have any of the following investigations for the previous hip problem? (tick those that apply)

X-Ray       Ultrasound scan       MRI scan       Other: (please describe) \_\_\_\_\_

Y N  
1 0

39. Have you ever had problems with the **opposite** hip?      Yes       No

Y N ?  
1 0 2

40. Has anyone in your immediate family had a history of hip pain/problems?

Yes       No       Don't know

P S A/U  
1 2 3  
Other=4

41. If 'Yes' to question 40, what is the relationship of this person to you? \_\_\_\_\_

No OAL Other na  
0 1 2 3 100

42. Do you know what the diagnosis or treatment was for this person? If yes, please give details below.

Participant ID \_\_\_\_\_

**Aggravating Positions and Activities:**

Do any of the following positions or activities **produce or aggravate** your **hip pain**?  
 If so, please circle the number that represents how much pain that position/activity causes.

	No Pain	1	2	3	4	5	6	7	8	9	Pain as Bad as it Can Be
43. Walking	0	1	2	3	4	5	6	7	8	9	10
44. Walking up or down stairs/slopes	0	1	2	3	4	5	6	7	8	9	10
45. Standing	0	1	2	3	4	5	6	7	8	9	10
46. Rising from sitting	0	1	2	3	4	5	6	7	8	9	10
47. Getting in/out of car	0	1	2	3	4	5	6	7	8	9	10
48. Putting on shoes/socks	0	1	2	3	4	5	6	7	8	9	10
49. Squatting	0	1	2	3	4	5	6	7	8	9	10
50. Driving	0	1	2	3	4	5	6	7	8	9	10
51. Jogging/running	0	1	2	3	4	5	6	7	8	9	10
52. Twisting activities	0	1	2	3	4	5	6	7	8	9	10
53. Other.....	0	1	2	3	4	5	6	7	8	9	10
54. Other.....	0	1	2	3	4	5	6	7	8	9	10

Office Use

**Activities that Relieve/Ease your Symptom(s)**

Y N  
1 0

55. Is there anything you can do that relieves or eases your hip pain significantly? Yes  No

56. If yes to Q 55, please describe what you can do to relieve or ease your pain:

0= Rest  
1=Meds  
2=Mobs  
3=Other

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**Pain Behaviour:**

Y N  
1 0

57. During the last 4 weeks, have you had hip pain (groin or upper thigh) on most days?  
 Yes  No

Y N  
1 0

58. During the last 4 weeks, have you had hip pain while walking downstairs or down slopes?  
 Yes  No

Y N  
1 0

59. During the last 4 weeks, have you noticed any limitation in the range of movement of one or both hips?  
 Yes  No

Participant ID \_\_\_\_\_

Y N 1 0	60. Is your hip usually stiff first thing in the morning?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
Mins (n) na=100	61. If "yes", <u>about</u> how long does it take for most of the hip stiffness to go?	____ (hrs) ____ (mins)	
Y N 1 0	62. Are you usually able to sleep on the side of your sore hip at night?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
Y N 1 0	63. Does sleeping on your <u>good</u> side usually cause pain in your sore hip?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
Y N 1 0	64. Do you wake with hip pain:	Yes <input type="checkbox"/>	No <input type="checkbox"/>
	if 'Yes'		
Mins (n) na=100	a. how long does it take for you get back to sleep?	____mins ____hours	
Y N na 1 0 100	b. do you <u>need</u> to get up to relieve your pain?	Yes <input type="checkbox"/>	No <input type="checkbox"/>

### Occupation & Leisure Details

Y N 1 0	65. Are you currently in paid employment?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
0=Sed 1=Activ 100=na	66. If 'Yes' to Q 65, what is your occupation: _____		
	67. If 'No' to Q 65 Please tick the one description below that best describes you:		
1 2 3 4 5 na=100	<input type="checkbox"/> Retired	<input type="checkbox"/> At home with children/maternity leave	<input type="checkbox"/> Unable to work due to illness/injury
	<input type="checkbox"/> Unemployed	<input type="checkbox"/> Other: _____	
Y N 1 0	68. Are you usually involved in any regular sport, recreational activities or hobbies?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
	If 'Yes', please describe: _____		
	69. If 'Yes' to Q68, please indicate the demands of your sport/recreational activity in terms of your hip:		
1 2 3 100	<input type="checkbox"/> Low hip demand (e.g. swimming, easy gardening, handcrafts)		
	<input type="checkbox"/> Moderate hip demand (e.g. walking, golf, moderate gardening, biking)		
	<input type="checkbox"/> High hip demand (e.g. running, hiking, racket sports, soccer, volleyball, contact sport, heavy landscaping, weight lifting)		
Y N na 1 0 100	70. Is your hip pain currently preventing you from participating in any of these activities?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
	If yes, which one(s)? _____		
Y N 1 0	71. Is there anything else that you would like to tell us about your hip problem? If so, please use the space below to do so:		

## Appendix 5. Consent form for interviews

<h3>Consent Form to Participate as a Person with a History of Hip Pain in an Interview</h3>	
---	---

Project title: **The reliability of clinical tests used in the assessment of hip joint pain**

Project Supervisor: **Professor Peter McNair**

Researcher: **Mr Steven White**

- I have read and understood the information provided about this research project in the information sheet dated 22 March 2010.
- I have had an opportunity to ask questions and have them answered.
- I am 20 years old or over.
- I have had a painful hip that has been assessed by a sports physician or orthopaedic surgeon.
- I am not currently having treatment for my hip.
- I have not had surgery on my hip joint.
- I have not consulted a medical or other health professional for low back pain within last 12 months.
- I do not have severe pain with my hip.
- I understand that my identity is confidential.
- I understand that notes will be taken during the interview and that it will also be audio-taped.
- I understand that I may withdraw myself or any information that I have provided for this project at any time prior to completion of data collection, without being disadvantaged in any way.

*Please Circle*

I consent to taking part in this research		Yes	No
I wish to receive a summary of the results		Yes	No
I consent to my interview being audio-taped and transcribed		Yes	No
I would like the audio-tape of my interview returned to me at the end of the study		Yes	No
Full Name			
Address			
Phone	H	W	Mob

Participant signature: ..... Date:.....

**Approved by the Auckland University of Technology Ethics Committee on 14 May 2010  
AUTEK Reference number 10/44**

## Consent Form to Participate as an Expert in an Interview



Project title: **The reliability of clinical tests used in the assessment of hip joint pain**

Project Supervisor: **Professor Peter McNair**

Researcher: **Mr Steven White**

- I have read and understood the information provided about this research project in the information sheet dated 22 March 2010.
- I have had an opportunity to ask questions and have them answered.
- I am 20 years old or older.
- I understand that my identity is confidential unless I choose it to be otherwise.
- I understand that notes will be taken during the interview and that it will also be audio-taped.
- I understand that I may withdraw myself or any information that I have provided for this project at any time prior to completion of data collection, without being disadvantaged in any way.

*Please Circle*

<b>I consent to taking part in this research</b>		Yes	No
I wish to receive a summary of the results		Yes	No
I consent to my interview being audio-taped and transcribed.		Yes	No
I would like the audio-tape of my interview returned to me at the end of the study		Yes	No
I wish to be identified as an 'expert' contributor to this project		Yes	No
Full Name			
Address			
Phone	H	W	Mob

Participant signature: ..... Date:.....

**Approved by the Auckland University of Technology Ethics Committee on 14 May 2010  
AUTEK Reference number 10/44**

## Appendix 6. Interview study information sheet

### Interview Participant Information Sheet



#### Date Information Sheet Produced:

22 March 2010

#### Project Title

The reliability of clinical tests used in the assessment of hip joint pain.

#### An Invitation

I am Steven White, a senior lecturer and PhD candidate at AUT University and a qualified, registered physiotherapist. You are invited to participate in a research study. This study will contribute to my PhD.

Participation is completely voluntary (your choice) and you may withdraw from the study at any time prior to the completion of data collection without giving a reason or being disadvantaged. Your participation in this study will be stopped should any harmful effects appear.

#### What is the purpose of this research?

The purpose of these interviews is to give the researcher a better understanding of hip joint pain prior to a follow-up study that aims to investigate the reliability of tests commonly used in a clinical setting to diagnose hip joint problems. If you agree, data collected from these interviews may also be used for further studies planned to follow on from the reliability study. Specifically, a study that determines how accurate hip joint tests are in the identification of pain that originates from within the hip and then another which will evaluate the effectiveness of physiotherapy treatment of hip joint pain.

You are either a person who has experienced such pain or a medical professional with experience in managing people with hip pain. The researcher wishes to talk to you about your respective experiences.

For those of you who have lived with hip pain, you will most likely have consulted various health professionals. If so, you have probably undergone various tests and procedures that were used to help the medical professional make a diagnosis in regard to your hip pain. You will have an opinion about this process and may well be able to provide insight that could help others who have yet to follow in your footsteps.

As a medical professional, you make decisions regarding the diagnosis and management of people with hip joint pain. You will have your own ideas regarding the value of information you gather from a person whether it be from the questions you ask, the tests you perform or investigations you have ordered. You may well have your own variations in the way you perform some tests or in how you interpret test results or other information you get from the person. These variations most likely reflect your own experiences and 'knowing' of what has previously helped you to make the most appropriate decisions for individuals.

The researcher wants to know more about what you think. Your thoughts will help the researcher to make sure that the next phase of this research is well informed.

#### Are you eligible to participate in this project?

You are eligible if you are either:

1. A clinician with expertise in the management of hip joint pain or
2. A person who has had a painful hip that has been assessed by a sports physician or orthopaedic surgeon, but who is not currently having treatment

However, unfortunately you are **not** eligible if you are a person with a history of hip pain and

1. You have had hip joint surgery or
2. You have consulted a medical or other health professional for low back pain within last 12 months or
3. You still have severe pain with your hip

#### **What will happen in this research?**

This study involves one session of approximately 30-60 minutes. It will be undertaken at a venue convenient to you. You will be met by the researcher and the interview process will be explained. Your informed consent to participate will be gained. The researcher will initiate a conversation with you. There will be some standard questions that he will ask however, it is intended that the interview is not structured in a manner that limits your ability to contribute in any way that you think is appropriate. You should feel comfortable bringing up any topic of conversation that you feel is relevant.

If you agree, the researcher will use a digital voice recorder to record the conversation so that he can reflect on what was discussed at a later date. You are free to change your mind about the recording of the conversation at any stage during the interview and to have the recording deleted if you so wish. You will not have to explain why you have changed your mind.

#### **What are the discomforts and risks?**

There is no physical discomfort or risk associated with the interview.

#### **How will these discomforts and risks be alleviated?**

You are free at any time to stop the interview or to refuse to answer any question. You will not be subjected to any physical tests or manoeuvres. You are welcome to have a chaperone or support person attend the interview with you. The researcher is a qualified, registered physiotherapist with 30 years experience. His conduct is subject to professional and statutory standards of ethical practice.

#### **What are the benefits?**

The results of this study may help to improve the diagnosis and management of people with hip joint pain. It may well lead to earlier diagnosis of hip problems and in doing so decrease the number of complications that result from delayed diagnosis. It may also decrease the need for more expensive and invasive testing procedures.

#### **What compensation is available for injury or negligence?**

In the unlikely event of a physical injury as a result of your participation in this study, rehabilitation and compensation for injury by accident may be available from the Accident Compensation Corporation, providing the incident details satisfy the requirements of the law and the Corporation's regulations.

#### **How will my privacy be protected?**

Any information we collect will not be able to be identified as belonging to you. All data collected will be coded so that you will only be identified by a number. The researchers will be the only people who have access to this information. All information will be kept in a secure room and in a locked filing cabinet for a 10-year period and then be destroyed. Electronic data will be stored on the researcher's personal laptop that requires password access that only the

researcher is privy to. All electronic data will be erased following standard procedures, 10 years after the completion of the study

**What are the costs of participating in this research?**

There is no financial cost to you to participate in this research.

**What opportunity do I have to consider this invitation?**

You have a week to decide whether you wish to take part in the study. You have a right to choose not to participate. If you agree to take part you are free to withdraw from the study at any time prior to the completion of data collection, without having to give a reason.

**How was I chosen to participate in this research?**

If you are a medical professional, you were identified through professional networks as someone with a reputation as an expert in the area of hip joint pain. If you are a person with a history of hip joint pain, you were identified through personal networks. Someone who knew you and about your hip pain, knew about the research and thought you would like to contribute.

**How do I agree to participate in this research?**

If you agree to participate in the study please complete the attached consent form.

**Will I receive feedback on the results of this research?**

If you wish to have a copy of the results of this research, please indicate this on the consent form. Results will be available after the study is completed and published.

**What do I do if I have concerns about this research?**

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor, Peter McNair, [peter.mcnair@aut.ac.nz](mailto:peter.mcnair@aut.ac.nz), and a 921 9999 ext 7143.

Concerns regarding the conduct of the research should be notified to the Executive Secretary, AUTEK, Madeline Banda, [madeline.banda@aut.ac.nz](mailto:madeline.banda@aut.ac.nz), 921 9999 ext 8044.

**Whom do I contact for further information about this research?**

Please feel free to contact the researcher if you have any questions about this study.

**Researcher Contact Details:**

Steve White; [steve.white@aut.ac.nz](mailto:steve.white@aut.ac.nz) ph 09 921 9999 ext 7073.

*Note: The participant should retain a copy of this form.*

This study has received approval from the AUT Ethics Committee on 14 May 2010  
Reference number: 10/44

## **Appendix 7. Operational definitions of tests and measures**

All tests investigated across both reliability and diagnostic accuracy studies are described below. Most tests were incorporated into all studies. However, some were not. Tests included in the reliability studies are denoted by the number 1 in superscript (<sup>1</sup>) beside the test name. Similarly, those included in the diagnostic accuracy study are denoted by the number 2 in superscript (<sup>2</sup>). Prior to the description of test performance, the following text provides detail common to each test, depending on whether or not it was investigated as a pain provocation test, a measure of range of movement (ROM) or as a strength test.

### **Pain Provocation Tests**

Prior to the physical examination, participants were introduced to the term ‘familiar pain’ i.e. pain in a similar site and of a similar nature to the pain that they typically experience with their troublesome hip during activities of daily living. With each test, participants were required to report the onset of any pain. The examiner then clarified if that pain was a ‘familiar pain’. The examiner also determined if this pain was felt in the primary site or a secondary site (as defined by the pain diagram that the participant had completed at baseline). Then the examiner asked for consent to gently apply further load until either end range of movement was achieved (as determined by significant tissue resistance) or until the participant’s pain response indicated that the application of further load would be inappropriate. The participant was then asked to rate the intensity of the pain provoked using the NPRS.

A ‘positive’ test for the painful hip was considered to be reproduction of ‘familiar’ pain provided the pain intensity was greater than or equal to 2-points on the NPRS. A positive test for the asymptomatic hip was considered to be pain greater than 2-points on the NPRS felt in the hip region (groin, greater trochanteric or deep buttock regions).

Strength (kg) and range of motion (degrees) measurements were obtained using a previously validated hand held dynamometer (HHD) that incorporates both a force transducer and gravity dependent inclinometer (Industrial Research Ltd; Christchurch; NZ). The HHD was calibrated before data collection.

### **ROM Tests**

To measure range of motion, the examiner first placed the HHD on the body part to be moved (in a standardised fashion) and ‘zeroed’ the device such that the initial start position was determined with reference to the vertical plane. Next, the body part was moved through to its end range where the final position was recorded. End range of movement was determined by the examiner feeling significant tissue resistance or by the patient stating that further motion would be “unacceptably uncomfortable/painful”. The HHD automatically subtracted the initial starting position from the final

position and displayed the actual range of motion. Three repetitions were performed with approximately 30 seconds between measurements.

To minimise any potential bias associated with these measures, the HHD was positioned in such a manner that the values could not be seen during the actual test manoeuvre. The device used ‘captures’ and retains the range of motion achieved during the test manoeuvre. This allows the examiner to remove the device from the patient, and therefore to read and record the value after the test has been completed.

### **Strength Tests**

For each strength test, the participant was instructed how to perform the test and then a sub-maximal practice test was performed. Next a ‘practice’ maximum voluntary contraction (MVC) was performed (the force produced was not measured). Finally, three repetitions of a MVC were performed with a 120 second rest between each repetition. During the measurement the examiner used a standardised instruction: “Go, push hard, push, push, push” to encourage the participant to produce a MVC and to make sure the contraction was maintained for a five second period. The contractions were all isometric ‘make’ force, performed against the HHD supported by the examiner. The patients were instructed to stabilise themselves by holding on to the testing plinth during the performance of each test. The force transducer measured the peak force that the participant generated during the five-second hold.

To minimise any potential bias associated with these measures, the HHD was positioned in such a manner that the values could not be seen during the actual test manoeuvre. The device used ‘captures’ and retains the highest peak force achieved during the test manoeuvre. This allows the examiner to remove the device from the patient, and therefore to read and record the value after the test has been completed.

### **Test Descriptions**

#### *Log Roll*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was positioned supine. The examiner passively rotated the patient’s leg back and forth so that the hip moved from internal to external rotation. The examiner used one hand on the mid-thigh area and the other at the ankle to apply this force. The knee was maintained in full extension. End ROM was avoided so that the rotation applied minimal stress to the soft-tissue structures surrounding the hip.

#### *Flexion Internal Rotation (FIR)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The hip was flexed to 90<sup>0</sup>, fully internally rotated and then slight overpressure added. No adduction or abduction was allowed.

*Flexion External Rotation (FER)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The painful hip was flexed to 90<sup>0</sup>, fully externally rotated and then slight overpressure added. No adduction or abduction was allowed.

*Full Flexion Internal Rotation (FFIR)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The examiner fully flexed the hip (in the sagittal plane), then moved the hip into full internal rotation and added slight overpressure.

*Full Flexion External Rotation (FFER)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The examiner fully flexed the hip (in the sagittal plane), then moved the hip into full external rotation and added slight overpressure.

*Impingement Test (FADDIR)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The hip was flexed to 90<sup>0</sup>, fully adducted and then fully internally rotated. Slight overpressure was applied in the direction of internal rotation whilst very subtle variations in the degree of flexion and adduction were explored.

*Flexion Adduction Compression (FADC)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The hip was flexed to 90<sup>0</sup> and adducted 20<sup>0</sup> and then axial compression was applied through the long axis of the femur.

*Compression in Internal Rotation @ 90<sup>0</sup> Flexion (FIRC)*<sup>1</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The examiner flexed the hip 90<sup>0</sup>, fully internally rotated the hip and then applied axial compression through the long axis of the femur.

*Quadrant Test*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The painful hip was passively flexed to 90° and then adducted until passive resistance to the movement was detected. Then, the examiner *gently* moved the hip through an arc of motion of approximately 90<sup>0</sup> degrees so that the hip finished in a fully flexed position without any adduction. Next, the hip was returned to the start position by reversing the direction of movement and following the same arc of motion.

*FABER*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was supine. The painful hip was moved to a position of flexion, abduction and external rotation by placing the lateral malleolus on the contralateral knee, just above the patella. The examiner stabilized the opposite anterior superior iliac spine (ASIS) whilst simultaneously applying slight overpressure to move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). Slight overpressure was then added.

*Bent Knee Fall Out (BKFO)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and ROM test. The participant was supine lying. The painful hip was flexed so that the medial malleolus of the ankle was aligned with the medial joint line of the opposite knee. Then the non-painful hip was flexed to match the position of the painful side. This manoeuvre ensured consistency in the start position of the painful hip in the sagittal plane. The HHD was positioned on the medial aspect of the femur of the painful hip, centred on the apex of the medial femoral condyle. In this position the HHD was zeroed. The examiner stabilized the ASIS on the non-painful hip side and the painful hip was slowly abducted/externally rotated. If pain was reproduced during this motion, the participant was asked if the examiner could move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). The participant was instructed to advise if further movement caused any increase in intensity of pain and to tell the examiner to stop if they felt uncomfortable about the level of discomfort. Range of movement was measured at the point that either pain or resistance limited further motion or when any compensatory motion of the pelvis was noted. Three repetitions were performed with ROM being recorded after each repetition.

After ROM measurements had been completed, the manoeuvre was repeated without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not and, if so, the patient was required to rate the intensity of that pain.

*Full Flexion (FF)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and ROM test. The participant was in crook lying. The hip to be measured was positioned in neutral abduction/adduction and rotation. The examiner palpated the participant's lumbo-sacral junction with his left hand whilst passively flexing the right hip through the sagittal plane with his right hand positioned on the upper tibia, close to the knee. At the point that the lumbar spine started to flex, the therapist encouraged the participant to gently resist this spinal movement and then the examiner continued to flex the hip. End range of movement was defined as that point where no further hip flexion could be gained without causing flexion of the lumbar spine. The examiner observed the motion, making sure that the hip stayed in neutral adduction/abduction and neutral rotation.

Once EROM was reached, the participant was instructed to hold the hip being tested at that exact point in the range with his hands clasped around his knees whilst the examiner positioned the HHD on the

anterior aspect of the thigh at a level previously marked (35mm proximal to the superior pole of the patella). The HHD was zeroed at this point and then the hip slowly returned to the start position. The HHD recorded the ROM travelled during the return to neutral. Three repetitions were performed with ROM being recorded after each repetition.

After ROM measurements had been completed, the manoeuvre was repeated without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not and, if so, the patient was required to rate the intensity of that pain.

#### *Resisted Abduction (RAB)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was lying supine on a plinth with the both hips in neutral abduction/adduction and knees extended. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test.

For measures of strength, the HHD was positioned on the apex of the lateral malleolus of the side to be tested. The examiner placed his other hand on the plinth beside the lateral aspect of the ankle on the side not being tested. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD. Simultaneously, the participant was required to exert pressure with the opposite leg against the examiners hand to enhance stability. Peak force generated during this test was recorded.

When employed as a pain provocation test, the HHD was not used and instead the examiner applied his hands directly to the participant's lateral malleolus. The participant was required to maintain the contraction for 5 seconds.

#### *Resisted Adduction (RAD)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was lying supine on a plinth with the both hips in neutral abduction/adduction and knees extended. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test.

For measures of strength, the HHD was positioned on the apex of the medial malleolus of the side to be tested. The examiner placed his other hand on the plinth beside the medial aspect of the ankle on the side not being tested. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD. Simultaneously, the participant was required to exert pressure with the opposite leg against the examiners hand to enhance stability. Peak force generated during this test was recorded.

When employed as a pain provocation test, the HHD was not used and instead the examiner applied his hands directly to the participant's medial malleolus. The participant was required to maintain the contraction for 5 seconds.

*Resisted Extension (RE)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was supine. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test. He was also told he could push down on to the plinth with the leg not being tested.

For measures of strength, the HHD was placed directly under the heel and the examiner held the heel 2 centimetres above the plinth. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD to the heel. The participant was required to keep the knee fully extended. Peak force generated during this test was recorded.

When employed as a pain provocation test, the HHD was not used and instead the examiner cradled the participant's heel directly. The participant was required to maintain the contraction for 5 seconds.

*Adduction in Standing (SAD)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was standing and asked to weight bear on the 'sore hip side' and to progressively drop the pelvis towards the opposite side so that adduction occurred on the 'sore hip side'.

*Internal Rotation in Standing (SIR)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was standing on one leg (the painful side) and positioned so that the pelvis was in neutral i.e no anterior or posterior rotation or hip adduction or abduction. The examiner provided support to the participant's pelvis to help with balance and to slowly move the participant's pelvis and trunk to create internal rotation at the hip joint. Gentle overpressure was added at the EROM. The participant was instructed to keep the knee of the weight bearing leg fully extended.

*External Rotation in Standing (SER)*<sup>1,2</sup>

This test was investigated purely as a pain provocation test. The participant was standing on one leg (the painful side) and positioned so that the pelvis was in neutral i.e no anterior or posterior rotation or hip adduction or abduction. The examiner provided support to the participant's pelvis to help with balance and to slowly move the participant's pelvis and trunk to create external rotation at the hip joint. Gentle overpressure was added at the EROM. The participant was instructed to keep the knee of the weight bearing leg fully extended.

*Squat to chair (Squat)*<sup>1</sup>

This test was investigated purely as a pain provocation test. The participant stood with the back of their knees light touching a chair (without armrests). The chair had been previously adjusted so that when the participant was seated, their feet rested flat on the floor and their knee joint was at 90<sup>0</sup> of flexion. Participants were instructed to slowly sit on the chair, without using their hands to lower themselves. This manoeuvre was demonstrated to the participant by the examiner first.

*Rise from chair (Rise)<sup>1</sup>*

This test was investigated purely as a pain provocation test. The participant started from a seated position with their feet in the same position as they were during the ‘Squat to Chair’ test (above). They were required to initiate the movement by leaning forward at the hip so that their head was positioned well forward of their knees. Participants were asked to rise from the chair without using their hands. This manoeuvre was demonstrated to the participant by the examiner first.

*Adduction in 4 point kneel (ADDKn)<sup>1</sup>*

This test was investigated purely as a pain provocation test. The participant adopted the 4-point kneeling position with their hands shoulder width apart and arms vertical. Initially, femurs were positioned vertically and the lumbar spine was positioned in maximum lumbar extension (induced by anteriorly tilting the pelvis). Tibias were positioned so that they were parallel to each other (i.e. neutral rotation) with the apex of the medial condyles of the femurs 100mm apart. The participant slowly shifted his weight towards the ipsilateral (painful) side maintaining the starting degree of hip flexion, full lumbar spine extension and neutral hip rotation.

*Flexion in 4-point kneel (4PtFlex)*

This test was investigated purely as a pain provocation test. The participant adopted the 4-point kneeling position with their hands shoulder width apart and arms vertical. Initially femurs were positioned vertically and the lumbar spine was positioned in maximum lumbar extension (induced by anteriorly tilting the pelvis). Tibias were positioned so that they were parallel to each other (i.e. neutral rotation) with the apex of the medial condyles of the femurs 100mm apart. Then, the participant flexed his hips by slowly sitting back towards his heels, maintaining lumbar spine extension and neutral hip rotation.

*Internal Rotation in Prone (IRPr)<sup>1,2</sup>*

This test was investigated as both a pain provocation test and ROM test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was prone with feet over the edge of the treatment table. The hip being measured was placed in 0° of abduction, and the contralateral hip in 30° of abduction. The reference knee was flexed to 90°. For measures of ROM, the HHD was positioned along the medial aspect of the distal shaft of the tibia (centred over the apex of the medial malleolus of the ankle) and zeroed whilst in this start position.

The tibia was passively moved in the frontal plane to produce hip internal rotation. The examiner monitored the tibio-femoral joint for any motion that could be construed as hip rotation. The examiner stabilised the participant’s pelvis by pushing on the sacrum with his other hand (in a manner that controlled any motion of the pelvis). The examiner also ensured that the femur did not abduct or adduct. If pain was reproduced during this motion, the participant was asked if the examiner could move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). The participant was instructed to advise if further movement caused any increase in

intensity of pain and to tell the examiner to stop if they felt uncomfortable about the level of discomfort. Range of movement was measured at the point that either pain or resistance limited further motion or when any compensatory motion was noted. Three repetitions were performed with ROM being recorded after each repetition.

When used as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not.

*External Rotation in Prone (ERPr)<sup>1,2</sup>*

This test was investigated as both a pain provocation test and ROM test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The hip being measured was placed in 0° of abduction, and the contralateral hip in about 30° of abduction. The reference knee was flexed to 90°. For measures of ROM, the HHD was positioned along the lateral aspect of the distal shaft of the tibia (centred over the apex of the lateral malleolus) and zeroed whilst in this start position.

The tibia was passively moved in the frontal plane to produce hip external rotation. The examiner monitored the tibio-femoral joint for any motion that could be construed as hip rotation. The examiner stabilised the participant's pelvis by pushing on the sacrum with his left hand (in a manner that controlled any motion of the pelvis). The examiner also ensured that the femur did not abduct or adduct. If pain was reproduced during this motion, the participant was asked if the examiner could move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). The participant was instructed to advise if further movement caused any increase in intensity of pain and to tell the examiner to stop if they felt uncomfortable about the level of discomfort. Range of movement was measured at the point that either pain or resistance limited further motion or when any compensatory motion was noted. Three repetitions were performed with ROM being recorded after each repetition.

When used as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not.

*Extension in prone (EPr)<sup>1,2</sup>*

This test was investigated purely as a pain provocation test. The participant was prone. The hip was maintained in neutral abduction/adduction and rotation. The hip was extended passively by the examiner, holding under the participant's femur (just above the patella). The participant's knee was kept in full extension. The examiner monitored the participants ipsilateral ASIS to ensure that it remained in contact with the table. The examiner observed the lumbar spine so that any compensation there was noticed and controlled.

*Internal rotation sitting (IRSit)<sup>1,2</sup>*

This test was investigated as both a pain provocation test and ROM test. The participant was sitting on a plinth with hips and knees flexed to 90<sup>0</sup> and the legs 'hanging' in a relaxed starting position. For measures of ROM, the HHD was positioned along the medial aspect of the distal shaft of the tibia (centred over the apex of the medial malleolus) and zeroed whilst in this start position.

The tibia was passively moved in the frontal plane to produce hip internal rotation. The examiner monitored the tibio-femoral joint for any motion that could be construed as hip rotation. The examiner also ensured the participant did not move his trunk or pelvis and that the femur did not abduct. If pain was reproduced during this motion, the participant was asked if the examiner could move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). The participant was instructed to advise if further movement caused any increase in intensity of pain and to tell the examiner to stop if they felt uncomfortable about the level of discomfort. Range of movement was measured at the point that either pain or resistance limited further motion or when any compensatory motion was noted. Three repetitions were performed with ROM being recorded after each repetition.

When used as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not.

*External rotation sitting (ERSit)<sup>1,2</sup>*

This test was investigated as both a pain provocation test and ROM test. The participant was sitting on a plinth with hips and knees flexed to 90<sup>0</sup> and the legs 'hanging' in a relaxed starting position. For measures of ROM, the HHD was positioned along the lateral aspect of the distal shaft of the tibia (centred over the apex of the lateral malleolus) and zeroed whilst in this start position.

The tibia was passively moved in the frontal plane to produce hip external rotation. The examiner monitored the tibio-femoral joint for any motion that could be construed as hip rotation. The examiner also ensured the participant did not move his trunk or pelvis and that the femur did not abduct. If pain was reproduced during this motion, the participant was asked if the examiner could move the hip further into the range until resistance to further motion occurred (i.e. end range of passive movement). The participant was instructed to advise if further movement caused any increase in intensity of pain and to tell the examiner to stop if they felt uncomfortable about the level of discomfort. Range of movement was measured at the point that either pain or resistance limited further motion or when any compensatory motion was noted. Three repetitions were performed with ROM being recorded after each repetition.

When used as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. Slight overpressure was added at the end of the ROM to determine if a familiar pain was provoked or not.

*Resisted flexion (RF)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was sitting on a plinth with hips and knees flexed to 90°. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test. For measures of strength, the HHD was positioned on the anterior aspect of the participant's thigh, five cms proximal to the patella. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD to the thigh. Peak force generated during this test was recorded. The test was repeated 3 times. When employed as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. The participant was required to maintain the contraction for 5 seconds.

*Resisted internal rotation (RIR)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was sitting on a plinth with hips and knees flexed to 90°. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test. The examiner placed one hand over the medial femoral condyle and instructed the participant to maintain contact against the hand to stop abduction occurring during the test. For measures of strength, the HHD was positioned on the apex of the lateral malleolus. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD to the lower leg. Peak force generated during this test was recorded. The test was repeated 3 times. When employed as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. The participant was required to maintain the contraction for 5 seconds.

*Resisted external rotation (RER)*<sup>1,2</sup>

This test was investigated as both a pain provocation test and strength test in the reliability studies, but only as a pain provocation test in the diagnostic accuracy study. The participant was sitting on a plinth with hips and knees flexed to 90°. The participant was instructed to hold on to the sides of the plinth so that he could stabilise himself during the test. The examiner placed one hand on the lateral aspect of the participant's knee and instructed the participant to maintain contact against the hand to stop adduction occurring during the test.

For measures of strength, the HHD was positioned on the apex of the medial malleolus. The participant was instructed to exert a maximum effort against the HHD whilst the examiner applied resistance through the HHD to the lower leg. Peak force generated during this test was recorded. The test was repeated 3 times.

When employed as a pain provocation test, this same manoeuvre was performed but without the application of the HHD. The participant was required to maintain the contraction for 5 seconds.

## Appendix 8. Between-session influencing factors

Trial # 3 Painful Hip L  R  Participant ID \_\_\_\_\_

Office Use

### Response to Previous Assessment:

Y N  
1 0

1. Was your hip pain aggravated by the assessment last week? Yes  No

If no, please go directly to question 4

0 1 2 3 4  
na=100

2. If yes, approximately how long were you sore for after the examination?  
1-2 Hours  Several Hours  Overnight  1-2 days  More than 3 days
3. If the increased pain lasted for more than 24 hours, please circle one number to indicate the level of that pain.

No Pain 0 1 2 3 4 5 6 7 8 9 10 Pain as Bad as it Can Be

### Medication Use:

Y N  
1 0

4. Have you changed anything about your medication use in the week since your assessment last week? Yes  No  If no, please go directly to question 6

Y N  
1 0

5. If yes, please tick the box(s) that apply  
 Usually I don't take medication but this week I did (please give detail regarding what you took)

Na=100

- I increased my pain relief  I decreased my pain relief  
 I increased my anti-inflammatorys  I decreased my anti-inflammatorys  
 I changed my medication because of the increased pain caused by the assessment  
 I changed my medication because of the increased pain caused by something I did during the week  
 Other...please give detail/explain

### Activity Levels:

Y N  
1 0

6. Have you changed anything about your level of physical activity in the week since your assessment last week? Yes  No   
If no, please go directly to question 8

Y N  
1 0

7. If yes, please tick the box(s) that apply:  
 I increased my activity level (if so, please provide detail below)  
 I decreased my activity level (if so, please provide detail below)

Trial # 3 Painful Hip L  R  Participant ID \_\_\_\_\_

**Pain Severity:**

When answering these questions, think only of the pain you are experiencing in relation to your hip.

Please circle one number for each of the following 4 questions:

On average, how bad has your pain been:

	No Pain	0	1	2	3	4	5	6	7	8	9	10	Pain as Bad as it Can Be
8. In the morning over the past <u>two days</u> ?		0	1	2	3	4	5	6	7	8	9	10	
9. In the afternoon over the past <u>two days</u> ?		0	1	2	3	4	5	6	7	8	9	10	
10. In the evening over the past <u>two days</u> ?		0	1	2	3	4	5	6	7	8	9	10	
11. With activity over the past <u>two days</u> ?		0	1	2	3	4	5	6	7	8	9	10	

Please circle one number for each of the following 2 questions:

12. What is the WORST/HIGHEST level of pain you have experienced at ANY time over the last week

No Pain	0	1	2	3	4	5	6	7	8	9	10	Pain as Bad as it Can Be
---------	---	---	---	---	---	---	---	---	---	---	----	--------------------------

13. What is the LOWEST level of pain you have experienced at ANY time over the last week

No Pain	0	1	2	3	4	5	6	7	8	9	10	Pain as Bad as it Can Be
---------	---	---	---	---	---	---	---	---	---	---	----	--------------------------

**Other Factors:**

There are a number of factors that can influence the level of pain we feel on a day to day basis including changes in the **weather** and **stress** (at home or work).

Y N  
1 0

Is there anything else that **you think** may have influenced **your pain** levels over the past week and that may have influenced the results of our tests today?

Please use the space below or over page to record your thoughts.

## Appendix 9. Ethical approval for diagnostic accuracy studies



### Northern X Regional Ethics Committee

Ministry of Health  
3<sup>rd</sup> Floor, Unisys Building  
650 Great South Road, Penrose  
Private Bag 92 522  
Wellesley Street, Auckland  
Phone (09) 580 9105  
Fax (09) 580 9001

Mr Steven G White  
8a Rawene Road  
Birkenhead 16 Selwyn Road  
Auckland 0626

Dear Steven

Re: Ethics ref: **NTX/11/07/066** (please quote in all correspondence)  
 Study title: Intra-articular pathology of the hip: the diagnostic accuracy of clinical examination  
 Investigators: Mr Steven G White (Principal), Professor Peter McNair, Associate Professor Wayne Hing, Dr Mark Laslett  
 Locality: Auckland University of Technology, Auckland Radiology Group

Thank you for your response and amendments requested by the Northern X Regional Ethics Committee. This study is now given ethical approval. A list of members of the Committee is attached.

#### Approved Documents

- Protocol number [version dated July 2011]
- Information sheet/Consent form version [1 dated 2 June 2011]
- Questionnaire version [undated received 23/6/2011]
- Baseline screening and collection sheet [version dated 14/6/2011]

This approval is valid until 30 September 2014, provided that Annual Progress Reports are submitted (see below).

#### Access to ACC

For the purposes of section 32 of the Accident Compensation Act 2001, the Committee is satisfied that this study is not being conducted principally for the benefit of the manufacturer or distributor of the medicine or item in respect of which the trial is being carried out. Participants injured as a result of treatment received in this trial will therefore be eligible to be considered for compensation in respect of those injuries under the ACC scheme.

#### Amendments and Protocol Deviations

All significant amendments to this proposal must receive prior approval from the Committee. Significant amendments include (but are not limited to) changes to:

- the researcher responsible for the conduct of the study at a study site
- the addition of an extra study site
- the design or duration of the study
- the method of recruitment
- information sheets and informed consent procedures.

Significant deviations from the approved protocol must be reported to the Committee as soon as possible.

Annual Progress Reports and Final Reports

The first Annual Progress Report for this study is due to the Committee by 9 August 2012. The Annual Report Form that should be used is available at [www.ethicscommittees.health.govt.nz](http://www.ethicscommittees.health.govt.nz). Please note that if you do not provide a progress report by this date, ethical approval may be withdrawn.

A Final Report is also required at the conclusion of the study. The Final Report Form is also available at [www.ethicscommittees.health.govt.nz](http://www.ethicscommittees.health.govt.nz).

Requirements for the Reporting of Serious Adverse Events (SAEs)

SAEs occurring in this study must be individually reported to the Committee within 7-15 days only where they:

- are *unexpected* because they are not outlined in the investigator's brochure, and
- are not defined study end-points (e.g. death or hospitalisation), and
- occur in patients located in New Zealand, and
- if the study involves blinding, result in a decision to break the study code.

There is no requirement for the individual reporting to ethics committees of SAEs that do not meet all of these criteria. However, if your study is overseen by a data monitoring committee, copies of its letters of recommendation to the Principal Investigator should be forwarded to the Committee as soon as possible.

Please see [www.ethicscommittees.health.govt.nz](http://www.ethicscommittees.health.govt.nz) for more information on the reporting of SAEs, and to download the SAE Report Form.

Statement of compliance

The committee is constituted in accordance with its Terms of Reference. It complies with the *Operational Standard for Ethics Committees* and the principles of international good clinical practice.

The committee is approved by the Health Research Council's Ethics Committee for the purposes of section 25(1)(c) of the Health Research Council Act 1990.

We wish you all the best with your study.

Yours sincerely



Cheh Chua-Ethics Committees  
Administrator  
Northern X Regional Ethics Committee

## Appendix 10. Screening form for diagnostic accuracy study

Participant Name \_\_\_\_\_

### DA Study Initial Recruitment & Screening Form

	YES	NO	Comments
Do you have hip pain mainly in the <u>groin</u> or deep, lower buttock region			
Has your pain present for minimum of <u>1 month</u> ?			
Are you between 18 & 80 years of age?			
Have you had any hip joint surgery ?			
Do you currently have low back pain?			
Have you had treatment for low back pain in the last 12 months?			
Do you have any pins & needles in your legs?			
Do you have any medical conditions (eg asthma, rheumatoid arthritis)			
Are you pregnant?			
Do you have severe pain with your hip?			
Are you claustrophobic?			

## Appendix 11. Diagnostic accuracy study information sheet

Appendix 2

# Participant Information Sheet



### Date Information Sheet Produced:

2 June 2011

### Project Title

Diagnosis of hip joint pathology

### An Invitation

I am Steven White, a senior lecturer and PhD candidate at AUT University and a qualified, registered physiotherapist. You are invited to participate in a research study. This study will contribute to my PhD. You can take your time in considering this invitation provided that if you do want to participate, you have informed the researcher before you make your appointment for the Magnetic Resonance Imaging Arthrogram (MRIa) that your doctor has recommended you have to help him diagnose the cause of your hip pain. You may have a friend, family or whānau support to help you understand the risks and/or benefits of this study and any other explanation you may require.

### Participation

Participation is completely voluntary (your choice). You do not have to take part in this study and if you choose not to take part, this will not affect any future care or treatment.

If you do agree to take part in the study, you are free to withdraw from the study at any time, without having to give a reason, and this will in no way affect your health care. Participation in this study will be stopped should any harmful effects appear or if your specialist or the researcher feels it is not in your best interest to continue. The study is limited to 48 participants. The first 48 eligible people who volunteer to participate will be included.

### How was I chosen to participate in this research?

You are a person with hip joint pain and you will have responded to one of the advertisements the researcher placed in waiting rooms of selected medical professionals or in a local paper. The adverts in the medical rooms were placed there because it was expected that people with hip pain might be more likely to see advertising and information about the study in that environment than elsewhere. Alternatively, your sports or orthopaedic specialist may have mentioned the study to you because they felt that you might benefit from being involved and that you were appropriate to participate.

### What is the purpose of this research?

This study aims to investigate the accuracy of information from tests performed and questions asked by medical professionals in clinical setting to diagnose hip joint problems. This will help physiotherapists, doctors and surgeons to improve their ability to determine the cause of pain felt in the hip region.

If you agree, data collected from this study may also be used for further studies (approved by an appropriate Ethics Committee) planned to follow on from the current study. No material that could personally identify you will be used in any reports on this study unless your personal approval is given for the dissemination of results to specific persons. For more information on privacy issues please see the section below titled "How will my privacy be protected?"

### Are you eligible to participate in this project?

You are eligible to participate in this study if:

- your sports or orthopaedic specialist has decided that you would benefit from having a magnetic resonance imaging arthrogram (MRIa) and local anaesthetic injection and
- you are between the age of 18 and 80 years old and

- you have pain mainly in the groin or deep buttock region on one side which has been present for a minimum of one month

You are **not** eligible to participate in this study if you:

- are unable to speak English
- have a history of hip joint surgery
- have consulted a medical or other health professional for treatment of low back pain within the 12 months prior to your current episode of 'hip' pain
- have evidence of nerve compression in your back or legs (e.g. pins & needles, muscle weakness)
- have any systemic disease or illness
- are pregnant
- have severe pain

#### **What will happen in this research?**

You will be examined by the researcher once before and once after undergoing the MRIa procedure. The first examination will take approximately 60 minutes and will be scheduled to occur on the same day that you have the MRIa. The second examination needs to be performed within 60 minutes of the MRIa, before the anaesthetic wears off. It will take approximately 30 minutes. Both examinations will be undertaken at the North Shore Campus, AUT University. The MRIa will be performed at the premises of Auckland Radiology (Wairau Road; Glenfield). Auckland Radiology will provide you with detail about this procedure once your appointment with them has been made.

A step by step description of what will happen is described below.

#### A: Subject Preparation for Examinations

You will be met at AUT by the researcher and the experimental procedure will be explained. Your informed consent will be gained. You will be required to complete standardised forms that give background detail about you such as age, occupation and levels of activity as well as detail specific to your hip pain. This will include information about the exact site of your pain, the intensity of your pain and how it effects your day to day activities. You do not have to answer all the questions.

You will be given an opportunity in private to change into your shorts and t-shirt (or similar) so that the researcher can test your hip in a manner that you feel safe and comfortable. If you wish to have a chaperone present during the testing you are welcome to bring someone along with you or to ask the researcher to arrange for an appropriate person for you.

#### B: Equipment

The researcher will use equipment that measures the range of movement of your hip and the amount of force you can generate during tests. This equipment is not invasive and will not cause any damage to you or your clothes.

#### C: Test procedures

The researcher will perform a number of tests on your hip that will determine the joint range of motion and hip muscle strength. Additionally, some tests are performed to see if they reproduce any discomfort or pain that you consider to be very similar to that which you normally get with your hip. These tests are tests that are commonly used by physiotherapists and doctors for patients with hip joint pain. It is very likely that you have had most of them performed on you already if you have seen a medical professional because of your hip pain.

The researcher will measure and record detail about range of movement, strength and reproduction of any discomfort or pain that you report. You will be asked to 'score' the level of any discomfort/pain that you feel, on a scale that has been tested and shown to be a valid measurement scale for pain.

Next, you will go to the Auckland Radiology rooms in Wairau Road to have the MRIa that your sports or orthopaedic specialist has recommended for you. After this procedure, you will return to AUT so that your hip can be re-examined by the researcher. This will enable the researcher to be able to determine if there is any change in the degree of pain or disability that you reported prior to the procedure. The researcher will then report the results of this examination back to your specialist so that he/she knows the effect of the anaesthetic.

**What are the discomforts and risks?**

There is a small risk of some soreness remaining in your hip after the completion of the physical examination. This is no different to how you might feel after going to your doctor or physiotherapist for an examination to get a diagnosis of your hip pain. It is not expected that you will have any significant soreness. Any soreness you may have is likely to settle on its own within 1-2 days.

There are some risks associated with the MRIa and local anaesthetic. These will be detailed by Auckland Radiology who will give you the opportunity to discuss any questions regarding this procedure, its benefits and its risks. All procedures used in this study are part of normal patient management. No additional dose or exposures will be used. The risks are those normally accepted as routine in typical clinical practice in New Zealand.

**How will these discomforts and risks be alleviated?**

The researcher has 30 years of experience as a physiotherapist specialising in treating musculoskeletal pain and is well able to recognise any signs that would suggest that you might have a hip that is likely to be aggravated by the test procedures. This is part of the reason why you are asked to complete the initial questionnaires mentioned above. There are a number of details that help to identify those who might experience this type of flare up in symptoms including:

- What, if any, reaction you have had to previous testing
- How you react to day to day activities and/or sporting/occupational activities
- How long any flare ups that you have experienced in the past have taken to settle.

Importantly, during the test sessions, should you feel any excessive discomfort, you can advise the researcher and he will stop the performance of that test. If you wish, you can be withdrawn from the rest of the examination. If you feel any discomfort following the study, you should inform the researcher who will provide you with advice and options regarding the management of any discomfort.

The MRIa will be performed by a radiologist at Auckland Radiology experienced with this procedure. Auckland Radiology will conduct a 'safety check' with you to identify any factors that would indicate that you should not undergo this procedure. Auckland Radiology is accredited by the National Radiation Laboratory and by IANZ (International Accreditation NZ). IANZ accreditation is an international process for assessing and recognising the technical competence and the effective quality processes of a professional service and its staff. It is granted only after an exacting assessment against international standards NZS/ISO/IEC 17025 for calibration laboratories and includes peer review.

**What are the benefits?**

The results of this study may help to improve the diagnosis and management of people with hip joint pain. It may well lead to earlier diagnosis of hip problems and in doing so decrease the number of complications that result from delayed diagnosis. It may also decrease the need for more expensive and invasive testing procedures.

For you personally, the opportunity to have a comprehensive and systematic examination of your hip just prior to and shortly after the MRIa improves the information gained from this procedure. Accurate assessment of any change in pain after the anaesthetic is a key factor in determining the cause of your hip pain and the subsequent management. Normally this process is compromised by the difficulties encountered in trying to co-ordinate the examination of your hip both before and then after the MRIa and the MRIa procedure itself. This would require your specialist to be available twice, several hours apart on the same day that you can schedule the MRIa. Being a participant in this study will allow a more accurate assessment of any change in your pain.

**What compensation is available for injury or negligence?**

In the unlikely event of a physical injury as a result of your participation in this study, rehabilitation and compensation for injury by accident may be available from the Accident Compensation Corporation, providing the incident details satisfy the requirements of the law and the Corporation's regulations.

**How will my privacy be protected?**

Any information we collect will not be able to be identified as belonging to you. All data collected will be coded so that you will only be identified by a number. The researchers will be the only people who have access to this information. All information will be kept in a secure room and in a locked filing cabinet for a 10 year period and then be destroyed. All electronic data will be kept on the researchers password protected laptop.

**What are the costs of participating in this research?**

There is no financial cost to you to participate in this research. It will take approximately 90 minutes of your time. You will receive a small monetary payment as an acknowledgement of your gift of time and to help compensate for any travel costs.

If you have private medical insurance, please check with your insurance company before agreeing to take part in the trial. You should do this to ensure that your participation will not affect your medical insurance.

**How do I agree to participate in this research?**

If you agree to participate in the study, please advise the researcher (see contact details below). He will answer any questions you may have and will co-ordinate your assessment of the study and your MRIa. You will need to complete the attached consent form and give it to the researcher.

**Will I receive feedback on the results of this research?**

If you wish to have a copy of the results of this research, please indicate this on the consent form. Results will be available after the study is completed and published. Because of the nature of the research & publication process, it is likely that results will not be available for 12 to 24 months after your participation. If you wish to discuss findings specific to you this may be arranged by appointment with the researcher.

**What do I do if I have concerns about this research?**

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor, Professor Peter McNair, [peter.mcnair@aot.ac.nz](mailto:peter.mcnair@aot.ac.nz); Ph 09 921 9999 ext 7143

If you wish to contact an independent health and disability advocate:  
Free phone: 0800 555 050  
Free fax: 0800 2 SUPPORT (0800 2787 7678)  
Email: [advocacy@hdc.org.nz](mailto:advocacy@hdc.org.nz)

**Statement of Approval**

This study has received ethical approval from the (insert name of committee) Ethics Committee, ethics reference number (insert ethics reference number).

**Whom do I contact for further information about this research?**

Please feel free to contact the researcher (Steve White) if you have any questions about this study.

Steve White; [steve.white@aot.ac.nz](mailto:steve.white@aot.ac.nz) ph 09 921 9999 ext 7073.

*Note: The Participant should retain a copy of this form.*

## Appendix 12. Consent form for diagnostic accuracy study

Appendix 4



# CONSENT FORM

Project title: **Diagnosis of hip joint pathology**  
 Project supervisor: **Professor Peter McNair**  
 Researcher: **Mr Steven White**

English	I wish to have an interpreter	Yes	No
Deaf	I wish to have a NZ sign language interpreter	Yes	No
Māori	E hiahia ana ahau ki tetahi kaiwhaka Māori/kaiwhaka pakeha korero	Ae	Kao
Cook Island Māori	Ka inangaro au i tetai tangata uri reo	Ae	Kare
Fijian	Au gadreva me dua e vakadewa vosa vei au	Io	Sega
Niuean	Fia manako au ke fakaaoga e taha tagata fakahokohoko kupu	E	Nakai
Sāmoan	Ou te mana'o ia i ai se fa'amatala upu	loe	Leai
Tokelaun	Ko au e fofou ki he tino ke fakaliliu te gagana Peletania ki na gagana o na motu o te Pahefika	loe	Leai
Tongan	Oku ou fiema'u ha fakatonulea	Io	Ikai

(Please circle)

I have read and understood the information sheet dated 2nd June 2011 for volunteers taking part in the study designed to determine the accuracy of the examination of the hip joint	Yes	No
I have had the opportunity to discuss the study and am satisfied with the answers I have been given	Yes	No
I have had the opportunity to use whānau support or a friend to help me ask questions and understand the study	Yes	No
I understand that taking part in this study is voluntary (my choice) and that I may withdraw from the study at any time, and this will in no way affect my future health care or continuing health care	Yes	No
I understand that my participation in this study is confidential and that no material that could identify me will be used in any reports on this study	Yes	No
I understand that the investigation will be stopped if it should appear harmful to me	Yes	No
I understand the compensation provisions for this study	Yes	No
I have had time to consider whether to take part in the study	Yes	No
I know who to contact if I have any side effects from the study	Yes	No
I know who to contact if I have any questions about the study in general	Yes	No
I wish to receive a summary of the research results	Yes	No
I am aged 20 years or over	Yes	No
I consent to the use of information about me, collected in this study, for follow up studies directly related to this research, providing they are given ethical approval by an appropriate Health & Disability Ethics committee	Yes	No
I give permission to be contacted regarding the possibility of participating in further research studies relevant to hip pain at the conclusion of this study	Yes	No

Version 1, 2<sup>nd</sup> June 2011

I..... hereby consent to take part in this study.

(Full Name)

Date:	<input type="text"/>
Signature:	<input type="text"/>
Full names of researchers:	Steven White & Peter McNair
Contact phone number for researchers:	Steven White 09 921 9999 ext 7073 Peter McNair 09 921 9999 ext 7143
Project explained by:	<input type="text"/>
Project role:	<input type="text"/>
Signature:	<input type="text"/>
Date:	<input type="text"/>

This study has received ethical approval from the (insert name of committee) Ethics Committee,  
ethics reference number (insert ethics reference number).

*Researcher Copy*

## Appendix 13. Medical screening questionnaire for diagnostic accuracy study

Steve White MHSc (Hons), Dip MT; Dip Public Health, MNZCP, Senior Lecturer

Participant ID

Date



# Medical Screening Questionnaire

### General Screening Questions

	YES	NO
Have you recently had a fever?	<input type="radio"/>	<input type="radio"/>
Have you recently taken antibiotics or other medicines for an infection?	<input type="radio"/>	<input type="radio"/>
Have you recently had surgery?	<input type="radio"/>	<input type="radio"/>
Have you recently had an injection into any joint?	<input type="radio"/>	<input type="radio"/>
Have you recently had a cut or scrape or open wound?	<input type="radio"/>	<input type="radio"/>
Have you been diagnosed with an immune-suppressive disorder?	<input type="radio"/>	<input type="radio"/>
Have you used IV drugs?	<input type="radio"/>	<input type="radio"/>
Do you have a history of cancer?	<input type="radio"/>	<input type="radio"/>
Have you recently lost weight for no apparent reason?	<input type="radio"/>	<input type="radio"/>
Have you recently felt unwell or unusually tired?	<input type="radio"/>	<input type="radio"/>
Have you been waking with pain at night that has not let you get back to sleep reasonably quickly?	<input type="radio"/>	<input type="radio"/>
Have you recently taken a long car, bus, plane or train ride?	<input type="radio"/>	<input type="radio"/>
Have you recently been bedridden for any reason?	<input type="radio"/>	<input type="radio"/>
Have you ever had surgery to your painful hip or back?	<input type="radio"/>	<input type="radio"/>
Do you have any numbness or tingling anywhere in your legs?	<input type="radio"/>	<input type="radio"/>
Have you ever taken any steroids?	<input type="radio"/>	<input type="radio"/>

Next Page

**Medications**

Please list ALL medications you are currently taking:

**General Health**Do you have any other medical conditions for which you receive treatment? Yes  No 

If yes, please describe that condition(s):

**Smoking History**

Cigarettes per day:

Non-smoker  1-10  11-20  21-30  30+ **Surgical History**

Please list any surgery for which you had a general anaesthetic:

Date:

Surgery:

  
  
**Allergies**

Please list all known allergies:

## Appendix 14. Lower limb tasks questionnaire (LLTQ)



Health and Rehabilitation Research  
Institute  
AUT University

### LOWER LIMB TASKS QUESTIONNAIRE ACTIVITIES OF DAILY LIVING SECTION

#### INSTRUCTIONS

Please rate your ability to do the following activities over the **past month** by clicking the circle below the appropriate response.

If you did not have the opportunity to perform an activity in the **past month**, please make your *best estimate* on which response would be the most accurate.

Please also rate how important each task is to you in your daily life according to the following scale:

- 1 = Not important
- 2 = Mildly important
- 3 = Moderately important
- 4 = Very important

Please answer all questions.

	UNABLE	SEVERE DIFFICULTY	MODERATE DIFFICULTY	MILD DIFFICULTY	NO DIFFICULTY	IMPORTANCE OF TASK
1. Walk for 10 minutes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
2. Walk up or down 10 steps (1 flight)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
3. Stand for 10 minutes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
4. Stand for a typical work day	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
5. Get on and off a bus	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
6. Get up from a lounge chair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
7. Push or pull a heavy shopping trolley	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
8. Get in and out of a car	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
9. Get out of bed in the morning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
10. Walk across a slope/uneven ground	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>

Enquiries concerning this questionnaire: Prof. Peter McNair, Health and Rehabilitation Research Institute, Auckland University of Technology, Private Bag 92006, Auckland, New Zealand. email: [peter.mcnair@aut.ac.nz](mailto:peter.mcnair@aut.ac.nz) Phone: 921-9999 Ext 714, Health and Rehabilitation Research Institute, AUT University



TOTAL



Health and Rehabilitation Research  
Institute  
AUT University



### LOWER LIMB TASKS QUESTIONNAIRE Recreational Activities Section

#### INSTRUCTIONS

Please rate your ability to do the following activities over the **past month** by clicking the circle below the appropriate response.

If you did not have the opportunity to perform an activity in the **past month**, please make your *best estimate* on which response would be the most accurate.

Please also rate how important each task is to you in your daily life according to the following scale:

- 1 = Not important
- 2 = Mildly important
- 3 = Moderately important
- 4 = Very important

Please answer all questions.

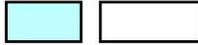
	UNABLE	SEVERE DIFFICULTY	MODERATE DIFFICULTY	MILD DIFFICULTY	NO DIFFICULTY	IMPORTANCE OF TASK
1. Jog for 10 minutes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
2. Pivot or twist quickly while walking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
3. Jump for distance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
4. Run fast/sprint	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
5. Stop and start moving quickly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
6. Jump upwards and land	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
7. Kick a ball hard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
8. Pivot or twist quickly while running	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
9. Kneel on both knees for 5 minutes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>
10. Squat to the ground/floor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	1 2 3 4 <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/>

Enquiries concerning this questionnaire: Prof. Peter McNair, Health and Rehabilitation Research Institute, Auckland University of Technology, Private Bag 92006, Auckland; New Zealand. email: [peter.mcnair@aut.ac.nz](mailto:peter.mcnair@aut.ac.nz) Phone: 921-9999 Ext 714, Health and Rehabilitation Research Institute, AUT University



TOTAL

## Appendix 15. Self-report Leeds assessment of neuropathic symptoms and signs (S-LANSS)



For these questions, think about how your pain has felt **over the last week**.

Please circle the descriptions that best match your pain. These descriptions may, or may not, match your pain no matter how it feels. Only circle the responses that describe your pain.

1. **In the area where you have pain, do you also have 'pins & needles', tingling or prickling sensations?**
  - a. No – I don't get these sensations  No
  - b. Yes – I get these sensations often  Yes
  
2. **Does the painful area change colour (perhaps look more mottled or more red) when the pain is particularly bad?**
  - a. No – The pain does not affect the colour of my skin  No
  - b. Yes – I have noticed that the pain does make my skin look different from normal  Yes
  
3. **Does your pain make the affected skin abnormally sensitive to touch? Getting unpleasant sensations or pain when lightly stroking the skin might describe this.**
  - a. No – The pain does not make my skin in that area abnormally sensitive to touch  No
  - b. Yes – My skin in that area is particularly sensitive to touch  Yes
  
4. **Does your pain come on suddenly and in bursts for no apparent reason when you are completely still? Words like 'electric shocks', jumping and bursting might describe this.**
  - a. No – My pain doesn't really feel like this  No
  - b. Yes – I get these sensations often  Yes
  
5. **In the area where you have pain, does your skin feel unusually hot like a burning pain?**
  - a. No - I don't have burning pain  No
  - b. Yes – I get burning pain often  Yes
  
6. **Gently rub the painful area with your index finger and then rub a non-painful area (for example, an area of skin further away or on the opposite side from the painful area). How does this rubbing feel in the painful area?**
  - a. The painful area feels no different from the non-painful area  No
  - b. I feel discomfort, like pins & needles, tingling or burning in the painful area that is different from the non-painful area  Yes
  
7. **Gently press on the painful area with your finger tip then gently press in the same way onto a non-painful area (the same non-painful area that you chose in the last question). How does this feel in the painful area?**
  - a. The painful area does not feel different from the non-painful area  No
  - b. I feel numbness or tenderness in the painful area that is different from the non-painful area  Yes

Total  /24

## Appendix 16. Baseline data collection form for diagnostic accuracy study

Steve White MHS (Hons), Dip MT; Dip Public Health, MNZCP, Senior Lecturer

Participant ID 

# BASELINE SCREENING & DATA COLLECTION

### Background Details

1. Name
2. Contact Phone: Home  Work  Mobile
3. Male  Female
4. Date of Birth (dd/mm/yyyy)
5. Ethnicity: NZ European  Maori  Pacific Islander  Asian  Other
6. Which is your painful hip? Left  Right
7. Which is your 'dominant' leg (for example, which leg would you normally use to kick a ball or push a spade into the ground)? Left  Right

### Irritability

8. Is it easy for you to aggravate your hip pain? Yes  No  Sometimes
9. If you aggravate your hip pain, how long does it normally take to settle again?  
Minutes  Hours  Overnight  1-2 days  3 or more days  Not Applicable
10. Do you take any medication for your hip pain? Yes  No  Sometimes
11. If yes (or sometimes), please provide detail of what you normally take and how often.

12. Are you sometimes totally free of pain, even if only for a few minutes? Yes  No

### Pain Severity

When answering these questions, think only of the pain you are experiencing in relation to your hip.

Use the scale below for each of the following 4 questions:

On average, how bad has your pain been: No Pain Pain as Bad as it Can Be

0 1 2 3 4 5 6 7 8 9 10

13. In the morning over the past two days?
14. In the afternoon over the past two days?
15. In the evening over the past two days?
16. With activity over the past two days?
17. Would you say that the severity of pain you have had over the last two days (as you have indicated in questions 13 to 16 above) is similar to what you 'typically' have over a two day period?  
Yes  No

Version 10/5/2012 Approved by the Northern X Regional Ethics Committee 9/8/2011 Study Ref MTX/11/07/066

[Previous Page](#)

[Next Page](#)

Steve White MHS (Hons), Dip MT; Dip Public Health, MNZCP, Senior Lecturer

Participant ID

18. If you answered 'No' to the question above (Q17), please select the number that best describes your AVERAGE level of your pain over a 'typical' two days.

	No Pain		Pain as Bad as it Can Be
<input style="width: 40px; height: 20px; border: 1px solid orange;" type="text"/>	0	1 2 3 4 5 6 7 8 9	10

Please select one number for each of the following 2 questions:

19. What is the **WORST/HIGHEST** level of pain you have experienced at ANY time over the last month

	No Pain		Pain as Bad as it Can Be
<input style="width: 40px; height: 20px; border: 1px solid orange;" type="text"/>	0	1 2 3 4 5 6 7 8 9	10

20. What is the **LOWEST** level of pain you have experienced at ANY time over the last month

	No Pain		Pain as Bad as it Can Be
<input style="width: 40px; height: 20px; border: 1px solid orange;" type="text"/>	0	1 2 3 4 5 6 7 8 9	10

21. How much has your hip problem affected your normal daily activity in the last week?  
(Please check the appropriate box)

1. <input type="checkbox"/>	2. <input type="checkbox"/>	3. <input type="checkbox"/>	4. <input type="checkbox"/>	5. <input type="checkbox"/>	6. <input type="checkbox"/>	7. <input type="checkbox"/>
No Disability	Almost no Disability	Minimal Disability	Some Disability	Moderate Disability	A lot of Disability	Maximum Disability

Steve White MHS (Hons), Dip MT; Dip Public Health, MNZCP, Senior Lecturer

Participant ID

### Type of Symptoms

22. What is the MAIN/WORST problem with your hip (tick only ONE):

Pain  Stiffness  Weakness  Leg giving way  Locking Other  (please describe):

23. How would you describe your MAIN pain?

Sharp pain  Aching (deep) pain  Other  (please describe):

24. Do you experience any of the following symptoms?

Crepitus (grinding/creaking or similar).	Yes <input type="radio"/>	No <input type="radio"/>
Clicking or clunking that is <u>painful</u> .	Yes <input type="radio"/>	No <input type="radio"/>
Clicking or clunking that is <u>not associated with pain</u> .	Yes <input type="radio"/>	No <input type="radio"/>
Locking of the hip.	Yes <input type="radio"/>	No <input type="radio"/>
Tingling/numbness anywhere in your legs.	Yes <input type="radio"/>	No <input type="radio"/>
Giving way of your leg because of pain or weakness.	Yes <input type="radio"/>	No <input type="radio"/>

### Location of Symptoms

Questions 25 and 26 refer to the pain drawing. See separate sheet.

**SCAN PAIN DRAWING****Previous Page****Next Page**

**Onset of Symptoms & Investigations**

27. How long have your current hip symptoms been present: (Please give the closest estimate for how many years/months or weeks)

 years months weeks

28. Describe how your current hip symptoms started:

Trauma   
e.g. fall, impact, accident

Strain/Injury   
e.g. strained lifting

Overuse   
e.g. repetitive action

Unknown

Briefly describe what you feel caused your current hip pain:

29. If your pain resulted from a specific event, how long after that event did you **FIRST** notice the pain?

Immediately

Within 48 hours

Longer than 48 hours

Not applicable

If the pain was not immediate, describe what you were doing when you **FIRST** noticed the pain:

30. From the time your symptoms **FIRST** started, when did you **FIRST** consult a medical professional (Dr or physio etc) about the problem?

Within 1 week

Within 1 month

Within 3 months

Within 6 months

Within 12 months

More than 12 months

Never

31. Have you already had any of the following investigations for this hip problem? (click all that apply)

X-Ray

Ultrasound scan

MRI scan

Other  (please describe)

### Aggravating Positions and Activities

Do any of the following positions or activities **produce or aggravate** your **hip pain**?

If so, please choose a number that represents how much pain that position/activity causes.

	No Pain	1	2	3	4	5	6	7	8	9	Pain as Bad as it Can Be	10
32. Walking												<input type="text"/>
33. Walking up <u>or</u> down stairs/slopes												<input type="text"/>
34. Standing												<input type="text"/>
35. Sitting												<input type="text"/>
36. Rising from sitting												<input type="text"/>
37. Getting in/out of car												<input type="text"/>
38. Putting on shoes/socks												<input type="text"/>
39. Squatting												<input type="text"/>
40. Driving												<input type="text"/>
41. Jogging/running												<input type="text"/>
42. Twisting activities												<input type="text"/>
43. Other <input type="text"/>												<input type="text"/>

### Activities that Relieve/Ease your Symptom(s)

44. Is there anything you can do that relieves or eases your hip pain significantly? Yes  No

45. If yes to Q 44, please describe what you can do to relieve or ease your pain:

### Pain Behaviour

46. During the last 4 weeks, have you had hip pain (groin or upper thigh) on most days?

Yes  No

47. During the last 4 weeks, have you had hip pain while walking downstairs or down slopes?

Yes  No

48. During the last 4 weeks, have you noticed any limitation in the range of movement of your hip?

Yes  No

49. Is your hip usually stiff first thing in the morning?

Yes  No

Steve White MHS (Hons), Dip MT; Dip Public Health, MNZCP, Senior Lecturer

Participant ID

50. If "yes" to Q 49, about how long does it take for most of the hip stiffness to go?  minutes51. Does it usually hurt to sleep on the side of your sore hip at night? Yes  No 52. Does it usually hurt your sore hip if you sleep on your good side? Yes  No 53. Do you wake with hip pain: Yes  No 

if 'Yes' to Q53

a. about how long does it take for you to get back to sleep?  mins  hoursb. do you need to get up to relieve your pain?  Yes  No**Occupation & Leisure Details**54. Are you currently in paid employment? Yes  No 

55. If 'Yes' to Q 54, what is your occupation:

56. If 'No' to Q 54, please select the one description below that best describes you: Retired  At home with children/maternity leave  Unable to work due to illness/injury Unemployed  Student  Other (please describe)57. Are you usually involved in any regular sport, recreational activities or hobbies? Yes  No 

If 'Yes', please describe:

58. If 'Yes' to Q57, please indicate the demands of your sport/recreational activity in terms of your hip:

 Low hip demand (e.g. swimming, easy gardening, handcrafts) Moderate hip demand (e.g. walking, golf, moderate gardening, biking) High hip demand (e.g. running, hiking, racket sports, soccer, volleyball, contact sport, heavy landscaping, weight lifting)59. Is your hip pain currently preventing you from participating in any of these activities? Yes  No 

If yes, which one(s)?

60. Is there anything else that you would like to tell us about your hip problem? If so, please use the space below to do so:

Version 10/5/2012 Approved by the Northern X Regional Ethics Committee 9/8/2011 Study Ref MTX/11/07/066

[Previous Page](#)[Next Page](#)

## Previous Hip Pain

61. Have you had previous problems with this (same) hip or groin? Yes  No  **If 'no' go to Q 69**

62. If yes, when did you FIRST have problems with this hip? (approximate date)

63. How did the previous problem first start?

Trauma  Strain/Injury  Overuse  Unknown   
 e.g. fall, impact, accident e.g. strained lifting e.g. repetitive action

Briefly describe what you feel caused your previous hip pain:

64. Did the problem fully resolve before the current episode? Yes  No

65. Did you consult a medical professional about the previous hip problem? Yes  No

66. Did they give you a diagnosis for that hip problem? Yes  No

67. If 'Yes' to Q66 please write down what the diagnosis was in the space below

68. Did you have any of the following investigations for the previous hip problem? (click those that apply)

X-Ray  Ultrasound scan  MRI scan  Other: (please describe)

69. Have you ever had problems with your other hip? Yes  No

70. If 'Yes' do you know what the diagnosis or treatment was for that hip? If yes, please give details below

71. Has anyone in your immediate family had a history of hip pain/problems?

Yes  No  Don't know

72. If 'Yes' to question 71, what is the relationship of this person to you?

73. Do you know what the diagnosis or treatment was for this person? If yes, please give details below.

SCAN PAGE

ATTACH FILE

SAVE WHOLE DOCUMENT

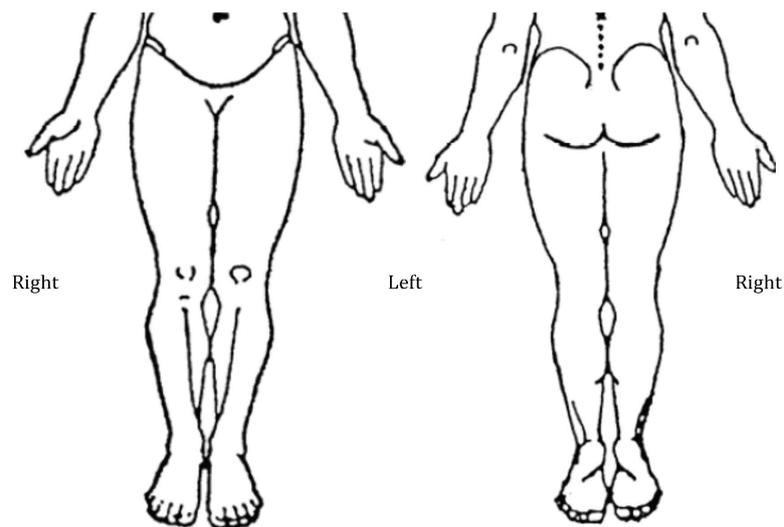
PRINT ALL

## Appendix 17. Body chart for diagnostic accuracy study

Participant ID

### Location of Symptoms

1. Please draw on the diagram below the areas where you feel your pain or other symptoms.
2. Use the colours in the 'Key' to describe the type of pain or symptom you have.
3. Please indicate with a cross (X) the location of your **main/worst** hip pain or symptom.



KEY	
Black	= aching pain
Red	= sharp or stabbing pain
Blue	= burning or coldness
Green	= pins & needles or numbness

### Appendix 18. Physical examination form for diagnostic accuracy study

Before FGAI or after  Painful Hip L  R  Participant ID  Date

No Pain 0 1 2 3 4 5 6 7 8 9 10 Pain as bad as it can be

#### Supine Tests

	Familiar Pain			Pain Score	ROM			
	Y	N	U		Good	Bad		
1. Log Roll	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
2. 90° FIR (90° F/EROM IR)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
3. 90° FER (90° F/EROM ER)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
4. FFIR (EROM F/EROM IR)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
5. FFER (EROM F/EROM ER)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
6. Impingement (90° F/AD/IR)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
7. 90° FAD Compression	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
8. Quadrant /Scour	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
9. FABERS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
10. FABERS (Modified)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
11. EROM Flexion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>	Mean FABERS G	<input type="text" value="0.0"/>	B	<input type="text" value="0.0"/>
12. ROM Flexion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
13. Resisted Abduction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>	Mean Flexion G	<input type="text" value="0.0"/>	B	<input type="text" value="0.0"/>
14. Resisted Adduction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
15. Resisted Extension	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				
16. De-rotation test @90	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>				

#### Prone Tests

17. IR EROM	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>
18. ER EROM	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>
19. Passive Extension	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>

Before FGAI or after  Painful Hip L  R  Participant ID

**Standing Tests**

	Familiar Pain			Pain Score
	Y	N	U	
20. Passive Adduction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21. EROM IR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. EROM ER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. Sustained 1 leg stand	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24. Height	<input type="text"/> metres			
25. Weight	<input type="text"/> kg			

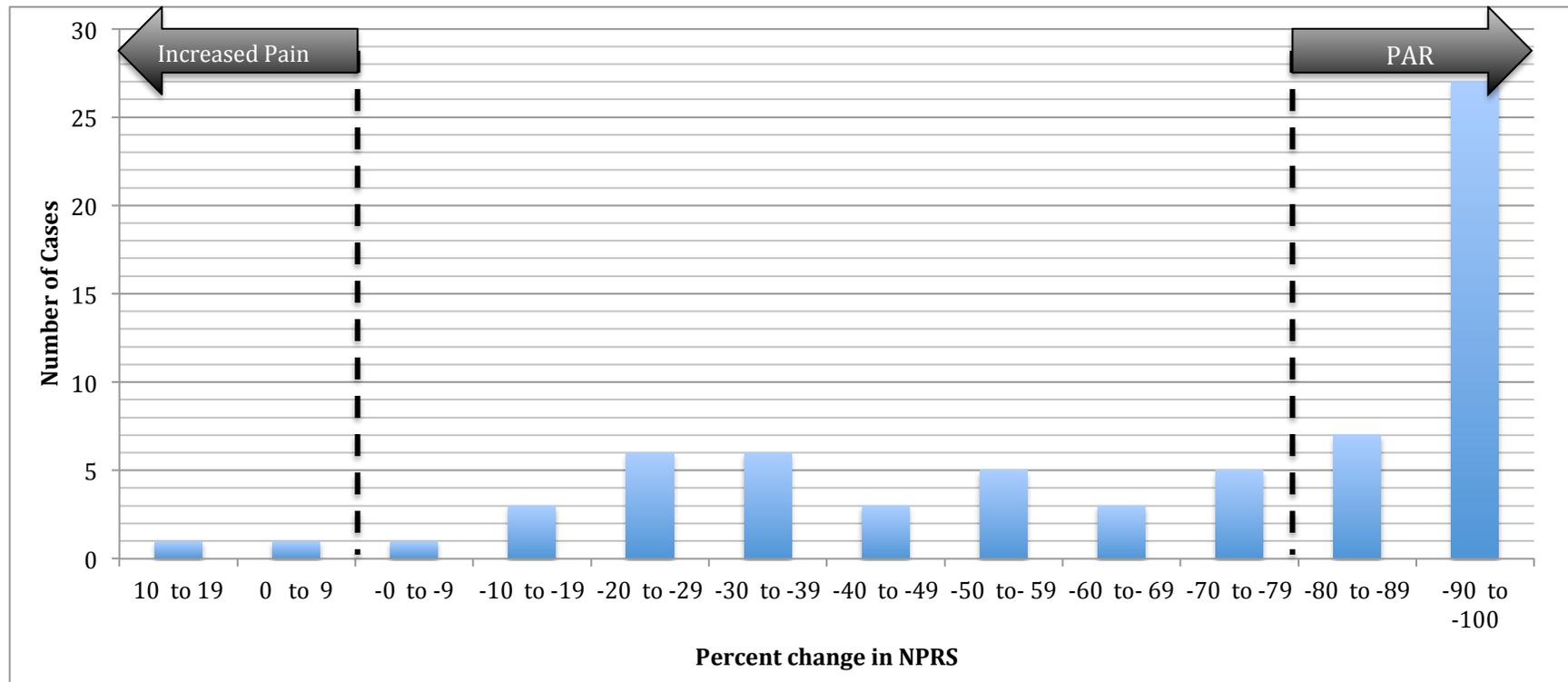
**4 Pt Kneeling**

26. Hip Flexion	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
-----------------	--------------------------	--------------------------	--------------------------	--------------------------

**Sitting**

	Y	N	U	Pain Score	ROM					
					Good			Bad		
27. IR ROM	<input type="checkbox"/>									
28. ER ROM	<input type="checkbox"/>									
29. Resisted Flexion	<input type="checkbox"/>									
30. Resisted IR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mean IR G	<input type="text"/> 0.0	Mean IR B	<input type="text"/> 0.0		
31. Resisted ER	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mean ER G	<input type="text"/> 0.0	Mean ER B	<input type="text"/> 0.0		
32. TOP Palpation Trochanter	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>						
33. Patient Specific Pain Provocation Manoeuvre # 1				<input type="checkbox"/>						
	<input type="text"/>									
34. Patient Specific Pain Provocation Manoeuvre # 2				<input type="checkbox"/>						
	<input type="text"/>									

### Appendix 19. Percent change in pain intensity following anaesthetic injection



NPRS, Numeric Pain Rating Scale; Positive values on 'Percent Change in NPRS' axis represent an increase in pain following anaesthetic injection; PAR, positive anaesthetic response.

**Appendix 20. Coordinates of the ROC curves for mean internal ROM**

Positive if Less Than or Equal To <sup>a</sup>	Sensitivity	1 - Specificity
14.000	.000	.000
16.000	.029	.000
18.333	.059	.000
19.833	.088	.000
20.333	.118	.000
21.167	.118	.029
22.667	.118	.059
24.333	.118	.088
25.167	.176	.088
25.667	.206	.088
26.167	.235	.088
26.500	.235	.118
27.833	.265	.118
29.167	.265	.147
30.333	.294	.147
31.500	.324	.147
32.000	.353	.147
32.500	.382	.147
32.667	.412	.147
32.833	.441	.147
33.333	.471	.206
33.667	.471	.265
33.833	.500	.265
34.167	.529	.265
34.500	.559	.324
35.167	.559	.353
36.500	.588	.353
37.333	.588	.382
37.833	.588	.441
38.333	.618	.441
38.667	.647	.471
39.500	.706	.500
40.167	.706	.529
40.333	.735	.529
40.500	.794	.529
40.833	.794	.559
41.167	.794	.647
41.667	.824	.647
42.500	.824	.676
43.500	.824	.706
44.333	.824	.735
44.833	.853	.735
45.167	.882	.735
46.500	.882	.765
48.000	.912	.794
48.667	.941	.824
49.167	.971	.824
50.167	.971	.853
52.667	.971	.882
55.167	.971	.912
56.167	1.000	.912
56.500	1.000	.941
66.667	1.000	.971
77.667	1.000	1.000

**Appendix 21. Coordinates of the ROC curves for the *difference* in range of movement between painful and non-painful hips**

Positive if Greater Than or Equal To <sup>a</sup>	Sensitivity	1 - Specificity	Positive if Greater Than or Equal To <sup>a</sup>	Sensitivity	1 - Specificity
-12.7000	1.000	1.000	-15.7000	1.000	1.000
-11.5000	1.000	.941	-12.0000	.971	1.000
-10.8000	1.000	.912	-9.1500	.971	.971
-9.8000	.971	.912	-8.5000	.971	.941
-9.0000	.971	.882	-7.1500	.971	.912
-7.8500	.971	.853	-6.1500	.971	.882
-6.8500	.971	.824	-5.8500	.971	.853
-6.5000	.941	.794	-5.5000	.941	.853
-6.1500	.912	.794	-5.0000	.941	.824
-5.3500	.912	.765	-4.0000	.912	.824
-4.3500	.882	.765	-3.0000	.882	.824
-3.8500	.853	.765	-2.5000	.853	.794
-3.5000	.824	.765	-2.1500	.794	.794
-3.1500	.794	.735	-1.8500	.765	.765
-2.5000	.794	.706	-1.5000	.765	.706
-1.6500	.706	.647	-1.1500	.765	.676
-1.1500	.706	.618	-.8500	.735	.588
-.5000	.706	.500	-.5000	.735	.529
.3500	.706	.471	-.1500	.706	.529
.8500	.706	.441	.1500	.676	.500
1.1500	.706	.412	.6500	.676	.471
1.5000	.706	.382	1.1500	.647	.441
1.8500	.676	.294	2.3000	.647	.412
2.1500	.647	.235	3.5000	.618	.412
2.6500	.559	.206	3.8500	.618	.382
3.1500	.529	.206	4.3500	.618	.294
3.5000	.500	.206	4.8500	.588	.294
3.8500	.471	.206	5.1500	.529	.265
4.1500	.441	.176	5.6500	.529	.206
4.5000	.441	.118	6.1500	.529	.176
4.8500	.441	.088	6.5000	.529	.118
5.1500	.412	.088	7.2000	.500	.118
5.8000	.382	.088	7.8500	.471	.118
6.5000	.353	.088	8.1500	.441	.118
7.2000	.324	.088	8.6500	.441	.088
7.8500	.294	.088	9.1500	.412	.088
8.5000	.265	.088	9.5000	.382	.088
9.3500	.235	.088	10.2000	.353	.088
10.2000	.206	.088	10.8500	.324	.088
10.8500	.176	.059	11.1500	.265	.088
11.1500	.176	.029	11.5000	.265	.059
12.3000	.147	.029	12.0000	.206	.059
13.8000	.118	.029	12.5000	.147	.029
14.6500	.059	.029	13.8500	.118	.029
16.0000	.029	.000	15.1500	.088	.029
18.0000	.000	.000	15.6500	.059	.029
			17.0000	.029	.029
			18.3500	.000	.029
			19.7000	.000	.000

**Internal Rotation ROM**

**BKFO ROM**

Note: Positive differences indicate that the asymptomatic side has a greater ROM than the symptomatic side

**Appendix 22. Coordinates of the ROC curves for age**

Positive if Greater Than or Equal To <sup>a</sup>	Sensitivity	1 - Specificity
15.000	1.000	1.000
16.500	1.000	.971
17.500	1.000	.941
18.500	1.000	.912
21.000	.941	.853
23.500	.941	.794
24.500	.941	.765
25.500	.882	.735
27.000	.882	.706
28.500	.882	.559
29.500	.853	.559
31.000	.824	.559
33.000	.794	.559
34.500	.765	.559
35.500	.735	.529
36.500	.735	.471
37.500	.735	.441
39.000	.706	.353
40.500	.647	.324
41.500	.588	.235
42.500	.559	.206
43.500	.471	.206
44.500	.471	.176
45.500	.412	.118
46.500	.353	.118
48.000	.324	.118
50.000	.235	.118
51.500	.206	.118
53.000	.147	.088
54.500	.088	.059
56.000	.059	.059
60.500	.029	.000
65.000	.000	.000

**Appendix 23. Coordinates of the ROC curves for mean BKFO ROM**

Positive if Less Than or Equal To <sup>a</sup>	Sensitivity	1 - Specificity
26.333	.000	.000
31.500	.029	.000
40.833	.059	.000
46.833	.088	.000
48.000	.088	.029
49.167	.088	.059
50.333	.118	.059
50.667	.147	.088
51.000	.176	.088
51.833	.206	.088
52.667	.235	.088
53.167	.265	.118
53.667	.265	.147
54.167	.294	.147
54.667	.294	.176
55.333	.324	.176
55.833	.412	.206
56.667	.441	.265
57.500	.441	.294
58.000	.441	.324
58.667	.471	.324
59.167	.500	.324
59.500	.529	.324
59.833	.559	.382
60.167	.559	.412
60.333	.588	.412
60.833	.618	.412
61.500	.647	.412
61.833	.647	.441
62.333	.676	.441
62.833	.676	.500
63.333	.676	.529
63.667	.735	.559
63.833	.765	.559
64.333	.765	.588
64.833	.765	.618
65.167	.765	.647
65.500	.765	.676
65.833	.794	.706
66.333	.794	.735
66.833	.824	.735
67.167	.853	.735
67.500	.853	.765
68.000	.882	.824
68.667	.882	.853
69.333	.882	.882
69.833	.912	.882
70.167	.941	.912
70.500	.971	.912
70.667	.971	.941
72.500	.971	.971
74.333	1.000	.971
75.333	1.000	1.000

## Appendix 24. Quadas 2 Critical Appraisal of DA studies for MR Imaging

Study	Risk of Bias				Applicability Concerns		
	Patient selection	Index Test	Reference test	Flow & Timing	Patient selection	Index Test	Reference test
Aprato	Low	Low	Unclear	Unclear	Low	Low	Low
Datir	Low	Low	High	Low	Low	Low	High
Devitt	Low	Low	Unclear	Unclear	Low	Low	Low
Reurink	Unclear	Low	Low	Unclear	Low	Low	Low
Sutter	Low	Low	High	High	Low	Low	Low

Low, low risk of bias or concern regarding applicability; High, high risk of bias or concern regarding applicability; Unclear, risk of bias or concern regarding applicability is unclear.

Appendix 25. MRA standardised reporting form



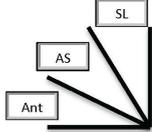
SPECIALIST RADIOLOGY + MRI  
GREENLANE

Participant #

NHI #

**Bone**

Bony Bump  No  Yes

Location  Head/Neck Offset  mm

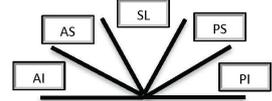
Alpha Angle Normal  No  Yes  Classical  degs Maximal  degs Position of Measurement

Osteophyte(s)  No  Yes Location

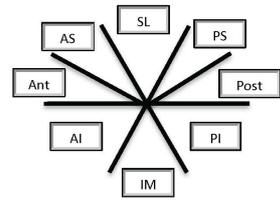
Size  Sml  Med  Lrg

Acetabular Version  Normal Anteversion  degs Retroversion  degs

Acetabular Cyst  No  Yes Location  Size  Sml  Med  Lrg



Femoral Cyst  No  Yes Location  Size  Sml  Med  Lrg



Oedema  No  Yes Location

Degree  Sml  Med  Lrg

Other bony abnormalities  No  Yes

Gad infusion into bone  No  Yes

Loose bodies  No  Yes Location

**Cartilage**

Normal  No  Yes

Thinning  No  Yes Location

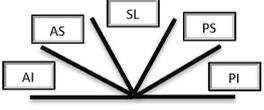
Degree  Mild  Mod  Severe

Fissuring  No  Yes Location

Degree  Mild  Mod  Severe

**Labrum**

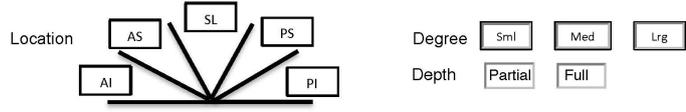
Normal  No  Yes

Fraying  No  Yes Location 

Degree  Minor  Mod  Extensive

Participant #   
NHI #

Tear  No  Yes Chondro-labral Separation  Linear  Complex  Notching



**Bursa**

Trochanteric Distended  No  Yes  Sml  Mod  Lrg  
Calcification  No  Yes

Iliopsoas Distended  No  Yes  Sml  Mod  Lrg  
Calcification  No  Yes

**Tendons**

Glut Min	Tendonopathy	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Min <input type="checkbox"/> Mod <input type="checkbox"/> Marked	Oedema	<input type="checkbox"/> No <input type="checkbox"/> Yes
	Tear	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		
Glut Med	Tendonopathy	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Min <input type="checkbox"/> Mod <input type="checkbox"/> Marked	Oedema	<input type="checkbox"/> No <input type="checkbox"/> Yes
	Tear	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		
Rec Femoris	Tendonopathy	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		
	Tear	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		
Iliopsoas	Tendonopathy	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		
	Tear	<input type="checkbox"/> No <input type="checkbox"/> Yes	<input type="checkbox"/> Sml <input type="checkbox"/> Mod <input type="checkbox"/> Lrg <input type="checkbox"/> Full		

**Ligamentum Teres**

Normal  No  Yes Swollen  Split  Ruptured

**Muscle**

Normal  No  Yes Findings

**Other Significant Pathology**

**XRay Findings**

Coxa Profunda/Overcoverage  No  Yes

Other:

Synovial Thickening  No  Yes

Intra-articular Pathology  No  Yes