NORTH-KARELIAN UNIVERSITY OF APPLIED SCIENCES
Physiotherapy degree program

Kivistö Heikki
Pasanen Tero

A VIDEO DEMONSTRATION OF THE 14 MOST VALID PHYSICAL EXAMINATION TESTS FOR THE SHOULDER

Thesis
October 2011

NORTH KARELIA
UNIVERSITY OF APPLIED SCIENCES

THESIS
October 2011
Degree Programme in Physiotherapy
Tikkarinne 9
FIN 80200 JOENSUU
FINLAND
Tel. (013) 260 6906

Author(s)
Heikki Kivistö, Tero Pasanen

Title
A video demonstration of the 14 most valid physical examination tests for the shoulder

Abstract

This thesis had two major purposes: (1) to review the literature for the most valid physical examination tests for the shoulder and (2) to produce a video demonstrating the correct performing techniques of the tests and how they should be interpreted.

Data for the literature review was collected from systematic reviews found from PubMed, Google Scholar and Cochrane databases. In these articles 14 clinical tests were found to be valid enough. Two of them are for impingement, six for full-thickness rotator cuff tears, three for anterior instability, one for labral tears and two for acromioclavicular joint pathologies. A 30–minute video was produced, demonstrating the performing techniques of the tests.

There are few high quality studies of the validity of the physical examination tests for the shoulder. The results of those studies suggest that most of the tests do not have consistent evidence of being acceptably valid for clinical purposes. However, the included tests seem to have value in affecting the likelihood of the condition of interest, especially when interpreted with a nomogram. Nonetheless, the results should be interpreted cautiously since the current evidence is inconsistent.

Our video provides an evidence based collection of the most valid physical examination tests for the shoulder with instructions of how to interpret the results. Therefore, it may help physiotherapists evaluate the underlying pathology with greater precision.

Language
English

Pages 38
Appendices 0
Pages of Appendices 0

Keywords
Shoulder, physical examination tests, clinical tests, orthopedic special tests, validity, instructional video.

POHJOIS-KARJALAN
AMMATTIKORKEAKOULU

OPINNÄYTETYÖ
Lokakuu 2011
Fysioterapian koulutusohjelma

Tikkarinne 9
80200 JOENSUU
p. (013) 260 6906

Tekijä(t)
Heikki Kivistö, Tero Pasanen

Nimeke
A video demonstration of the 14 most valid physical examination tests for the shoulder

Tiivistelmä

Opinnäytetyömme kaksi keskeisintä tavoitetta olivat: (1) selvittää kirjallisuuskatsauksen perusteella valideimmat kliiniset testit olkapäälle sekä (2) tehdä video, jolla esitetään kyseisten testien oikeat suoritustavat ja tulosten tulkinta.

Kirjallisuuskatsaukseen sisällytettiin systemaattiset katsaukset olkapään kliinisistä testeistä. Tietokantahakuina käytettiin PubMediä, Google Scholaria ja Cochranea. 14 testillä oli katsausten mukaan riittävä validiteetti ja tutkimusnäyttö. Kyseisistä testeistä kaksi on tarkoitettu ahdasolkaoireyhtymään, kuusi kiertäjäkalvosimen repeämälle, kolme anterioriselle instabiliteetille, yksi labrumin repeämälle ja kaksi AC–nivelen patologiselle tilalle. 30 minuutin video tehtiin sisältäen edellä mainitut testit.

Olkapään kliinisisten testien validiteetista on pieni määrä korkealaatuisia tutkimuksia. Niiden tulokset ovat pääasiassa epäjohdonmukaisia. Osa testeistä vaikuttaa kuitenkin olevan käyttökelpoisia muuttamaan arvioitavan patologisen tilan todennäköisyyttä erityisesti nomogrammia apuna käyttäen. Testien tuloksia täytyy kuitenkin tulkita varauksella, koska tutkimustulokset niiden validiteeteista eivät ole johdonmukaisia.

Videomme sisältää näyttöön perustuvan kokoelman valideimmista olkapään kliinisistä testeistä ja ohjeistuksen niiden tulosten tulkinnasta. Tuotoksemme voi auttaa fysioterapeutteja arvioimaan olkapään patologisia tiloja tarkemmin.

Kieli
Englanti

Sivuja 38
Liitteet 0
Liitesivumäärä 0

Asiasanat
Olkapää, olkanivel, AC–nivel, kliiniset testit, validiteetti, video.

TABLE OF CONTENTS

1 INTRODUCTION

Shoulder pain affects approximately one fourth of the population but diagnosing the source of the pain correctly remains challenging. The complex structure of the shoulder girdle and possible referred pain from the surrounding structures increase the risk of misdiagnosis. (Michener, Walsworth, & Burnet 2004; Parsons, Breen, Foster, Letley, Pincus, Vogel & Underwood 2007; Bongers 2001; Sizer, Phelps, Gilbert 2003; McFarland et al. 2010). However, making the correct diagnosis is crucial for maximizing the results and cost–effectiveness of the treatment. (Khan & Chien 2001; Rothstein 2001.)

In clinical setting, the most practical and economical way of narrowing down the diagnosis is to perform physical examination tests. These tests may sometimes be referred to as orthopedic special tests (OST) or clinical tests. In the physical examination tests for the shoulder the patient's arm is moved in a specific manner and then asked if the maneuver caused pain. There are over a hundred tests designed to diagnose shoulder pain but their usefulness has been questioned in recent reviews (Hegedus, Goode, Campbell, Morin, Tamaddoli, Moorman & Cook 2008; Hughes, Taylor & Green 2008; McFarland et al. 2010; Calvert, Chambers, Regan, Hawkins & Leith 2009; Munro & Healy 2009; Powell & Huijbregts 2006). Although six different systematic reviews have evaluated the physical examination tests for the shoulder, according to The Cochrane Library only the reviews by Hegedus et al. (2008) and Hughes et al. (2008) are of sufficient quality (Centre for Reviews and Dissemination 2011a; Centre for Reviews and Dissemination 2011b). Despite the current evidence, clinicians still seem to be performing tests that have little value in making a diagnosis even though tests with higher validity are available. The objective of our thesis was to help bridge this knowledge gap between researchers and clinicians. We created a video demonstrating the performing techniques for the most valid tests and presenting how the findings of the tests should be interpreted.

Because of the great variation in how the tests are performed in the medical literature, we hypothesized that images and text may not be the ideal media for

demonstrating the techniques accurately. Therefore, we opted to use video as our media to give clinicians a greater understanding of the position of the patient's arm as well as the applied strength, speed and direction of the movement. The thesis was designed with physiotherapists in mind who already have an understanding of the basic terminology, anatomy and pathophysiology of the shoulder girdle. Additionally, unlike others, we used the variations of the test techniques that had been evaluated in the included studies in Hegedus et al.'s review instead of the specific technique described by the originator of the test (Moen, Jan de Vos, Ellenbecker & Weir 2010; Tennent, Beach & Meyers 2003). By doing this, we ensured the validity figures match the techniques presented on the video.

Other video products similar to ours have been published recently (Cooke & Hegedus 2008; Hutchinson 2011). However, none of the videos include all the most valid OSTs for the shoulder. Furthermore, in the video by Hutchinson (2011), some invalid tests are demonstrated. The previous videos also haven't dealt with the interpretation of the test results in great detail. Because most OSTs cannot confirm a diagnosis when used individually and some tests are more useful than others, we introduce the use of a nomogram. With nomogram, the probability of the diagnosis being correct can be determined quickly and presented as a percentage value. This tool is introduced in chapter 4.4. Also, it enables stacking of the results of multiple tests, making OSTs significantly more valuable in clinical practice.

2 HOW DOES THIS WORK BENEFIT THE FIELD OF PHYSIOTHERAPY?

There seems to be great variation in what physical examination tests physiotherapists use when examining a painful shoulder. Also, the performing technique and the interpretation of the tests vary in the scientific literature. For instance, the Apprehension test has been performed either standing or supine, bilaterally or unilaterally and with the positive criterion being either pain or apprehension.

Nonetheless, in the light of current evidence it should be performed standing, simultaneously on both arms and with positive criterion being apprehension. (Hegedus et al. 2008; Farber, Castillo, Clough, Bahk & McFarland 2006; Lo, Nonweiler, Woolfrey, Litchfield & Kirkley 2004.) Therefore, this video production may help physiotherapists and other medical professionals by showing which tests are best suited for confirming or ruling out common shoulder girdle pathologies. Even learning that most tests can only be used to either rule in or rule out a condition may be new to many physiotherapists. In addition, our work presents the most valid performing techniques for the tests to ensure the accuracy of the findings.

The correct interpretation of the findings is also crucial for the applicability of the tests. Our video production clearly defines the positive criteria for the tests as well as the magnitude of the tests' impact on the likelihood of the suspected pathological conditions. Further, presenting the use of a nomogram helps physiotherapists in quantifying the impact of the tests. A nomogram may also be used to strengthen the overall impact of the tests when more than one test is available for the same purpose. We also present a table by McGee (2002) for quantifying the impact of the tests quicker or if a nomogram is unavailable.

The international classification of functioning (ICF) helps to clarify the aspect of physiotherapy that this thesis aims to improve. From the viewpoint of the ICF, the physical examination tests for the shoulder are used to identify disorders in the body structures. Specifically, the structure of the shoulder region (ICF code s720) (World Health Organization 2011). Therefore, the physical examination tests only indicate the underlying pathology but cannot tell how the patient's body functions or activities and participation may be affected by the condition. These matters still need to be evaluated separately.

Furthermore, assessing a painful shoulder usually involves patient history taking, observation and examination. In the examination, joint range of motion, joint play and reflexes are evaluated. The area is palpated and OSTs or diagnostic imaging may be performed. It is important to keep in mind that the physical ex-

amination tests are only a part of the examination process and should not be emphasized over the other aspects. (Magee 2006, 207–308.)

3 UNDERSTANDING STATISTICAL INDICES

3.1 Prevalence

Prevalence describes the proportion of a given population that suffers from a certain condition at a particular time (Porta 2008, 105). For instance, according to Auge & Fischer (1998) the prevalence of atraumatic osteolysis in weightlifters is 27%. This means that 27% of the weightlifters that were included in the study group suffered from this condition at that time. The figure was then generalized to apply to the whole weightlifting population to help clinicians estimate how many weightlifters in their practice might suffer from this disorder. Nonetheless, the prevalence value should be interpreted cautiously as there may be considerable variation depending on the geographical location and the clinic. Regardless of the margin of error, having an estimation of prevalence of the disorder is crucial when performing OSTs for the shoulder (Agoritsas, Courvoisier, Combescure, Deom & Perneger 2010; Davidson 2002).

3.2 Validity – does the test measure what it was intended to?

Validity is a statistical indicator that describes how accurately a certain test measures that which it was intended to measure (Fletcher & Fletcher 2005, 19). For instance, a subacromial impingement test should produce a positive test result only when impingement truly is present and a negative finding only when it is not present. Further, the presence of any other shoulder girdle lesion should not interfere with the finding by producing a false positive result.

However, completely valid clinical tests for shoulder girdle pathologies do not exist, making false negative or false positive test results possible. This is why

the validity of each test needs to be evaluated to see which tests are able to give clinically relevant results. The most common quantitative measures for assessing the validity of a test are sensitivity, specificity, predictive values and likelihood ratios (Riddle & Stratford 1999; Silman & Macfarlane 2002, 112–113).

Diagnostic sensitivity means the probability of a patient who has a specific disease getting diagnosed correctly. Diagnostic specificity, on the other hand, indicates the probability of a patient without a specific disease getting diagnosed correctly. (Fritz & Wainner 2001; Riddle & Stratford 1999.) Optimally, the tests should be 100% sensitive and 100% specific. Conversely, a value of 50% would indicate a similar probability as tossing a coin, making the test useless. Often, clinical tests for the shoulder have either high sensitivity or high specificity but not both. They can still be useful, however. If a test is highly specific (> 98%), a positive test result confirms the presence of pathology. If a test is highly sensitive (> 98%), a negative test result rules the disorder out. (Davidson 2002; Riddle & Stratford 1999.)

Positive (PPV) and negative predictive values (NPV) are another way of measuring validity. This index attempts to refine the estimation of probability by including the prevalence of the lesion in the measure. However, because of the location and clinic specific quality of prevalence, predictive values are not the preferred method for evaluating the validity of clinical tests for the shoulder. (Fritz &Wainner 2001; Davidson 2002.)

The most suitable validity indices for clinical decision making are the likelihood ratios which are derived from sensitivity and specificity values. However, likelihood ratios do not include the prevalence of the lesion. By excluding prevalence, they are not subject to the problem caused by clinic and region specific prevalence. However, the clinician may still choose to include prevalence or any other information in the form of pre-test probability. This process is described further in chapter 4. (Davidson 2002; Deeks & Altman 2004; Riddle & Stratford 1999.)

Likelihood ratios are not expressed as percentage values like other validity indices but in decimal numbers. Value of 1 signifies no change in probability of

the patient having the condition of interest whereas values above 1 increase the likelihood and values below 1 lessen the likelihood. While all values away from 1 can be useful, OSTs with values above 10 or below 0.1 are considered to be valid enough to be diagnostic. (Davidson 2002; Deeks & Altman 2004.)

In the case of positive finding in a clinical test, positive likelihood ratio (+LR) is applied. In simple terms, it indicates how many times more likely it is for the patient who has the condition of interest to get a positive test result compared with those who don't have it. (Deeks & Altman 2004.) As an example, the Bear–hug test has a +LR of 7.5 (Hegedus et al. 2008). A patient with a positive Bear–hug test is approximately seven times more likely to have a subscapularis tear than a patient with a negative bear–hug test result. To get a more accurate interpretation of what the test result indicates, a nomogram is used. The use of nomogram is described in chapter 4.4. (Davidson 2002; Deeks & Altman 2004.)

With negative likelihood ratio (-LR), a similar simple interpretation is not possible. Instead, it must be used with pre-test probability. (Davidson 2002.) As an example, if the clinician has estimated the chance of a patient having a subscapularis tear to be 60%, a negative result in the Bear–hug test would reduce this probability to 32%. The process of how to interpret the test results is discussed in chapter 4.

3.3 Reliability – will the result change if the test is repeated?

Reliability is a statistical measure that describes the amount of measurement error present in the test result or in other words, how consistent the results of a test are when repeated. Low reliability may, for instance, be due to lack of clinician's proficiency, nonstandardization of the test or lack of patient's comprehension. While reliability is helpful in determining the usefulness of a clinical test, it is less important than validity. Sometimes, tests with high validity can still be useful irrespective of poor reliability. (Campbell, Machin & Walters 2007, 202–203; Gadotti, Vieira & Magee 2006; Rothstein 2001; Wainner 2003.)

There are three types of reliability but with OSTs for the shoulder, only the ones that measure how consistently clinicians are able to perform tests are relevant. That is, intra- and inter-rater reliability. Intra-rater reliability describes the similarity between test results when carried out by the same clinician and inter-rater reliability describes the similarity of test results when carried out by two or more clinicians. (Weir 2005; Gadotti et al. 2006; Marx, Menezes, Horovitz, Jones & Warren 2003.)

In the studies reviewed in this text, reliability of clinical tests has been quantified using intraclass correlation coefficient (ICC) and kappa coefficient (k). Intraclass correlation coefficient (ICC) measures test-retest, intra- and inter-rater reliability. It is commonly reported as a decimal number between 0 and 1, with 0 signifying no reliability and 1 indicating perfect reliability. Compared with a simple percent calculation of reliability, ICC also evaluates the effect of systematic error (e.g. patient being fatigued when performing a strength test) and random error (e.g. chance) on the reliability score. There are roughly 10 different ways to calculate ICC and the choice of calculation method depends on the study setting. (Weir 2005; Shrout & Fleiss 1979; Gadotti et al. 2006). Weir (2005) cautions against setting universal standards for interpreting ICC. However, to give a general idea of the interpretation, commonly accepted guidelines are presented in table 1 (Portney & Watkins 2008, according to Van der El 2010, 6).

Table 1. Guidelines for interpreting intraclass correlation coefficient (ICC).

ICC value	**Interpretation**
<0.75	Poor to moderate reliability
>0.75	Good reliability
>0.90	Reasonable reliability for clinical measures

Kappa coefficient (κ), reported as a decimal number between 0 and 1, is used to measure intra- and inter-rater reliability. Kappa of 1 indicates perfect reliability but surprisingly, kappa of 0 doesn't signify complete lack of reliability. It merely suggests that the examiners did not agree any more than would be expected to happen by chance alone. A negative kappa value is possible when the examiners have agreed less than would be expected by chance. Compared with a simple percent calculation of reliability, kappa corrects the figure by considering agreement caused by chance as well as the prevalence of the condition of in-

terest. Inclusion of prevalence has been criticized, however. As with predictive values, the kappa value is useful only for clinics that have the same prevalence of the condition as in the study in which the reliability was evaluated. Moreover, the reliability of rare conditions is likely to be estimated incorrectly. Thus, kappa coefficient should be interpreted with caution. Although the most commonly used guidelines for interpreting kappa coefficient have been criticized for being arbitrary, they are presented in table 2 for general reference (Viera & Garrett 2005; Sim & Wright 2005; Stemler 2004; Portney & Watkins 2008, according to Van der El 2010, 6).

Table 2. Guidelines for interpreting kappa coefficient (κ).

Kappa value	Interpretation
0.0–0.4	Poor to fair agreement
0.4–0.6	Moderate agreement
0.6–0.8	Substantial agreement
>0.8	Excellent agreement

4 HOW TO APPLY AND INTERPRET CLINICAL TESTS

4.1 Overview

The clinical examination of the shoulder usually begins with history taking. Based on the acquired information and the known prevalence of the disorder, the clinician should roughly estimate the probability of the patient having the suspected condition. This rough percentage estimate (e.g. 60 %) that is made prior to performing any OSTs is called the pre-test probability. If this estimation is neither very low nor very high, further testing is warranted to confirm the diagnosis (Davidson 2002).

After estimating the pre-test probability, the clinician should try to rule this condition in or out by performing OSTs. After performing the appropriate OSTs and interpreting their effect using a nomogram, the clinician has the final percentage estimate of the likelihood of patient having the condition of interest. This esti-

mate is called the post-test probability. The use of a nomogram is instructed in chapter 4.4.

Unfortunately, none of the OSTs that have been studied so far are valid enough to be diagnostic. Nonetheless, they do serve as a practical and inexpensive method of significantly influencing the clinician's estimation of the presence of the condition of interest. At present, diagnostic tests that are more time-consuming and expensive, such as MRI or arthroscopy are needed to make a reliable diagnosis.

4.2 Step 1. Estimating the probability of the condition prior to testing

Pre-test probability, expressed as a percentage value, is determined by the clinician prior to performing any clinical tests. The estimation is based on the prevalence of the condition, the clinician's subjective impression or a combination of both. (Davidson 2002; Riddle et al. 1999; Agoritsas et al. 2010.) Unfortunately, no clear guidelines exist for determining the pre-test probability accurately and the estimation may differ significantly between clinicians (Attia, Nair, Sibbritt, Ewald, Paget, Wellard, Patterson & Heller 2004). Careful assessment of the pre-test probability is important, however, because of the overall low likelihood ratios of the OSTs for the shoulder.

To illustrate this process, let's consider an example patient, a 25-year-old ice hockey player complaining of anterior shoulder pain. The patient experiences pain with weight–bearing, abduction of the shoulder and when lying on the affected side. The clinician considers rotator cuff tear and acromioclavicular (AC) joint pathologies to be the most likely causes of the pain. Considering the prevalence for a full-thickness rotator cuff tear is 41% and for AC joint pathology 45%, the clinician estimates the pre-test probability to be 50% for rotator cuff tear and 70% for AC joint pathology. (Reilly et al. 2006; Powell et al. 2006; Maritz & Oosthuizen 2002.)

4.3 Step 2. Performing the clinical tests

Some OSTs are better at ruling in and others are better at ruling out a condition. Therefore, the clinician needs to select the most suitable OSTs for the situation. If the pre-test probability is high, all OSTs for that condition with a high positive likelihood ratio (+LR) should be performed. If, on the other hand, the pre-test probability is low, OSTs for that condition with a low negative likelihood ratio (-LR) should be used. The clinician might be tempted to perform all the OSTs for the shoulder. However, because of the low validity of the OSTs, the estimation of pre-test probability and selectively heightening or lowering the probability is needed. (Davidson 2002.)

In the example case, Active Compression test would be performed to rule in AC joint pathology and External Rotation Lag Sign, Hawkins-Kennedy and Supine Impingement tests would be used to rule out a rotator cuff tear (Hegedus et al. 2008).

4.4 Step 3. Interpreting the test results

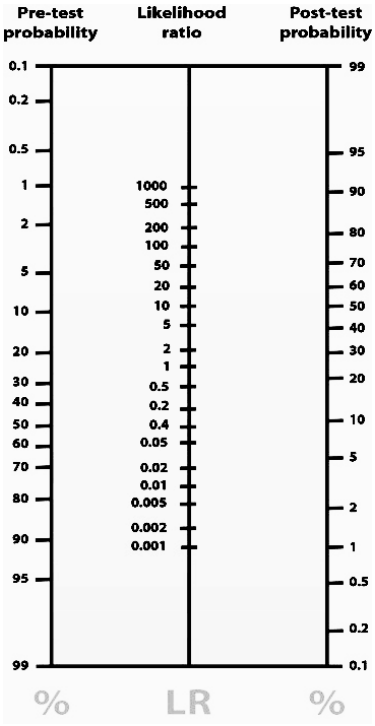

Figure 1. Fagan's likelihood nomogram.

A likelihood ratio nomogram (figure 1) is a tool that can be used for quickly calculation of post-test probability. We are aware of two different nomograms that can be used with OSTs for the shoulder: one by Fagan (1975) and one by Simon (2002). Both nomograms may be printed out on regular paper from the internet addresses found in the references section (Fagan 1975; Simon 2002).

With Fagan's nomogram the pre-test probability is marked with a pen on the leftmost vertical line. Then the +LR (in case of a positive test result) or -LR (in case of a negative test result) of the OST that was used is marked on the middle vertical line. Finally, a line is drawn through both marks until it reaches the rightmost vertical line. The post-test probability is found in this intersection. Multiple OSTs may be used in succession with the post-test probability of the last test becoming the pre-test probability for the following test. Performing multiple tests consecutively in this manner creates a larger shift in the probability of the condition being present. (Davidson 2002; Riddle 1999.)

With Simon's nomogram, the insert on the inside is slid until the correct pre-test probability reading lines up with the appropriate likelihood ratio reading. The post-test probability can then be read from the window at the top of the nomogram. (Simon 2002.)

Table 3. A table for interpreting likelihood ratios without a nomogram.

LR	**Approximate change in probability**
0.1	−45 %
0.2	−30 %
0.3	−25 %
0.4	−20 %
0.5	−15 %
2	+15 %
3	+20 %
4	+25 %
5	+30 %
6	+35 %
8	+40 %
10	+45 %

If a nomogram is unavailable, an alternative method may be used. A table developed by McGee (2002) (table 3) simplifies the process by presenting approx-

imations of the impacts of likelihood ratio values. For instance, according to the table, the Hawkins-Kennedy test (-LR 0.33) would reduce the likelihood of a rotator cuff tear by approximately 25%. However, the drawback of this method is inaccuracy, especially when the pre-test probability is close to the extremes, <10% or >90%. According to McGee (2002), the margin of error is between 4–10%. For instance in the following example, the Hawkins-Kennedy test lessens the probability only by 15%, giving an error margin of 10%. This source of inaccuracy is important to keep in mind when applying this method. (McGee 2002.)

Let's assume that in the example case, the Active Compression test (+LR 8.2) was positive and External Rotation Lag Sign (-LR 0.32) and Hawkins-Kennedy test (-LR 0.33) were negative. Using the nomogram as instructed above, the positive Active Compression test would raise the post-test probability for AC joint to 95%. The rule-out tests for rotator cuff tear should be interpreted consecutively. A negative External Rotation Lag Sign would give a post-test probability of 24%. Using this as the new pre-test probability, a negative Hawkins-Kennedy test would lower the post-test probability for rotator cuff tear to only 9%. In this example, the presence of AC joint pathology would have been confirmed (with post-test probability of 95%) and full-thickness rotator cuff tear ruled out (with post-test probability of 9%).

5 THE MOST VALID CLINICAL TESTS FOR THE SHOULDER

5.1 Subacromial impingement syndrome of any stage

Table 4. Statistical values for tests of any stage of SAIS.

Test	+LR	-LR	Reliability
Infraspinatus Muscle Strength test	4.2	0.64	n/a
Hawkins-Kennedy	2.12	0.33	k = 0.29

Subacromial impingement syndrome (SAIS) was first classified by Charles Neer into three stages by the severity of the condition. The first stage is characterized by subacromial edema and hemorrhage, the second stage by partial-thickness

rotator cuff tears, fibrosis or thickening and the third stage by full-thickness tears of the rotator cuff (FTT) or bony changes. Therefore, it is important to keep in mind that the diagnosis of SAIS of any stage could also present as a rotator cuff tear. (Neer 1983.)

The prevalence of SAIS in symptomatic shoulders is estimated to be 44–65% when all stages of SAIS are included. For SAIS of any stage, the Infraspinatus Muscle Strength test (+LR 4.2) may be used as a confirmatory test and the Hawkins-Kennedy test (-LR 0.33) may be performed as a rule-out test (Hegedus 2008). According to May, Chance-Larsen, Littlewood, Lomas & Saad (2010), the inter-examiner reliability for the Hawkins-Kennedy test is fair (k = 0.29). Reliability of the Infraspinatus Muscle Strength test has not been studied.

In the Infraspinatus Muscle Strength test the patient's shoulders are adducted to the side and the elbows are brought to 90 degrees of flexion. The patient is instructed to resist as the examiner applies an internal rotation force. The test is considered positive if the shoulder gives way because of weakness or pain or if External Rotation Lag Sign (ERLS) is positive. In the ERLS, the shoulder is brought to full external rotation and the patient is instructed to maintain this position. The test is considered positive if the patient is unable to hold this position unsupported. (Park, Yokota, Gill, El Rassi & McFarland 2005.)

To perform the Hawkins-Kennedy test, the arm of the involved side is brought to 90 degrees of shoulder flexion and 90 degrees of elbow flexion. The shoulder is passively internally rotated until either pain is elicited or the scapula begins to rotate. The positive criterion for this test is pain. (Park et al. 2005.)

5.2 Full-thickness rotator cuff tears

Table 5. Statistical values for tests of full-thickness rotator cuff tears.

Test	+LR	-LR	Reliability
Combined Hawkins-Kennedy, Painful Arc Sign and Infraspinatus Muscle Strength test	16.35	0.69	n/a
Palpation	29.91	0.04	n/a
Drop Arm	2.79	0.74	k = 0.28–0.66
Empty Can	2.99	0.58	k = 0.44–0.49
Belly Press	19.05	0.61	k = 0.31–0.65
Lift Off	Infinite	0.82	k = 0.28–0.30

Rotator cuff consists of the following muscles: infraspinatus, supraspinatus, subscapularis and teres minor (Lippert 2006, 110). The simplest way of classifying rotator cuff tears is to divide them into partial-thickness tears and full-thickness tears. Full-thickness tears go all the way through the rotator cuff tendon or tendons whereas partial-thickness tears, though may be wide, aren't deep enough to penetrate the muscle. (Habermeyer, Magosch, & Lichtenberg 2006, 20–21).

A rotator cuff tear can be seen as a part of normal aging process and it is often asymptomatic (Tempelhof, Rupp & Seil 1999; Worland, Lee, Orozco, SozaRex & Keenan 2003). For instance, Worland et al. (2003) showed that more than 40% of people over 50 years have a full-thickness rotator cuff tear (FTT) that is asymptomatic. Because of the degenerative nature of rotator cuff tears, their prevalence increases substantially after the age of 50 and continues to increase after that (Tempelhof et al. 1999; Moosmayer, Smith, Tariq & Larmo 2009; Ozaki, Fujimoto, Nakagawa, Masuhara & Tamai 1988).

The prevalence of FTTs in symptomatic shoulders is estimated to be 40.8%, but as was stated before it is highly age–dependent (Reilly et al. 2006). Palpation (+LR 29.91, -LR 0.04) is diagnostic for FTT. A combination of Hawkins-Kennedy, Painful Arc Sign and Infraspinatus Muscle Strength test (+LR 16.35, -LR 0.69) may be used to rule in the condition. Drop Arm (+LR 2.79, -LR 0.74) and Empty Can (+LR 2.99, -LR 0.58) may be used to further increase the likelihood of FTT. Belly Press (+LR 19.05, -LR 0.61) and Lift Off (+LR infinite, -LR 0.82) tests serve as rule-in tests for a full-thickness subscapularis tear. (Hughes

et al. 2008.) Drop Arm (k=0.28–0.66) and Belly Press (k=0.31–0.65) have fair to substantial, Empty Can (k = 0.44–0.49) moderate and Lift Off (k = 0.28–0.30) fair to moderate inter-examiner reliability. (May et al. 2010). There are no high quality studies with information about the inter-examiner reliability of palpation or the combination test.

Painful Arc Sign is used in combination with Infraspinatus Muscle Strength test and Hawkins-Kennedy to rule in a full-thickness rotator cuff tear. Hawkins-Kennedy and external rotation strength test are described in the previous chapter. The Painful Arc Sign is performed with the patient standing. The patient is asked to abduct the shoulder as far as it goes and then slowly lower the arm in the same arc of movement. The test is considered positive if the patient has pain or painful catching in the shoulder between 60 degrees and 120 degrees of shoulder abduction. (Park et al. 2005.)

Palpation of the rotator cuff tendons is performed with the patient standing. The arm of the involved side is brought to 90 degrees of elbow flexion. The clinician holds the proximal side of the patient's forearm with one hand and the other hand is used to palpate the insertion of the rotator cuff muscles. The involved arm is then brought in and out of external rotation of the shoulder while simultaneously extending the shoulder and bringing it back to neutral position. A positive finding is a rent felt in the tendons of the rotator cuff muscles. (Wolf & Agrawal 2001.)

The Drop Arm test is performed with the patient standing. The patient is asked to fully abduct the shoulder and then slowly lower the arm in the same arc of movement. The test is considered positive if the arm drops despite the patient's efforts to lower it slowly or if the patient has severe pain. (Park et al. 2005.)

In the Empty Can test the patient's arm is in 90 degrees of shoulder flexion and in neutral or full internal rotation. The patient is then asked to resist as the examiner places a downward force to the distal arm. The test is considered positive if the patient gives way. (Park et al. 2005.)

The Belly Press test is performed with the patient standing. The patient is instructed to place one hand on the abdomen and press against the belly. The test is performed on both sides. The test is considered positive if the involved side shows weakness compared with the unaffected side or the shoulder moves posteriorly. (Barth, Burkhart & De Beer 2006.)

In the Lift Off test the patient's hand is behind the back. The patient is then instructed to lift the hand off the back by internally rotating the shoulder. The test is considered positive if the patient is unable to perform the maneuver or if the maneuver is performed by extending the shoulder or the elbow. (Barth et al. 2006.)

5.3 Anterior instability

Table 6. Statistical values for instability tests.

Test	+LR	-LR	Reliability
Apprehension	20.2	0.29	ICC = 0.47
Relocation	10.4	0.20	ICC = 0.71
Anterior Release test	58.6	0.37	ICC = 0.63

Instability of the shoulder means inability of the humeral head to stay centered on the glenoid fossa. One way of categorizing different forms of instability is to divide them into three separate subcategories; traumatic structural, atraumatic structural and habitual nonstructural. Instability is also characterized by direction. It may be anterior, posterior or multidirectional. (Lewisa, Kitamurab, & Bayley 2004.) The following tests are for anterior instability and they cannot differentiate traumatic from atraumatic or structural from nonstructural instabilities (Farber et al. 2006; Lo et al. 2004).

The prevalence of instability is estimated to be 1.7% in the entire population (Kuhn 2010). The Apprehension test (+LR 20.2, -LR 0.29), Relocation test (+LR 10.4, -LR 0.20) and Anterior Release test (+LR 58.6, -LR 0.37) are all diagnostic for anterior instability (Hegedus et al. 2008). The inter-examiner reliability is moderate (ICC 0.47–0.63) for all the three tests (May et al. 2010).

The Apprehension test is performed standing. Both of the patient's arms are brought to 90 degrees of shoulder abduction and 90 degrees of elbow flexion. The shoulder is externally rotated until apprehension elicited or full external rotation is reached. The test is considered positive if apprehension is elicited. Pain is not considered a positive finding. (Farber et al. 2006.)

The Relocation test is performed with the patient supine. Patient's arms are brought to 90 degrees of shoulder abduction and 90 degrees of elbow flexion. The shoulder is externally rotated until apprehension is elicited or full external rotation is reached. In case of apprehension, a posteriorly directed force is applied to the humeral head. If this maneuver reduces the sensation of apprehension, the test is considered positive. (Farber et al. 2006.)

The Anterior Release test is performed with the patient supine. The posteriorly directed force applied in the Relocation test is suddenly released. The test is considered positive if this maneuver causes apprehension. (Lo et al. 2004.)

5.4 SLAP tears

Table 7. Statistical values for SLAP tear tests.

Test	+LR	-LR	Reliability
Biceps Load Test II	30	0.10	k = 0.815

Glenoid labrum is a fibrocartilage located around glenoid fossa. Its role is to make the glenoid fossa deeper and more stable. (Lippert 2006, 111.) SLAP (superior labral anterior to posterior) tear is a superior tear of the glenoid labrum that runs from posterior to anterior direction. The tear may extend to the attachment of the biceps tendon. (Snyder, Karzel, Del Pizzo, Ferkel & Friedman 1990.) The prevalence of SLAP lesions in symptomatic shoulders is estimated to be 26% (Kim, Queale, Cosgarea & McFarland 2003). The Biceps Load II test (+LR 30, -LR 0.10) may be used to both confirm and rule out a SLAP lesion (Hegedus et al. 2008). The inter-examiner reliability for the test is reported to be excellent (k = 0.815) (Kim, Ha, Ahn, Kim & Choi 2001).

The Biceps Load Test II is performed with the patient supine. The shoulder of the involved side is brought to 120 degrees of abduction and 90 degrees of external rotation. The elbow is brought to 90 degrees of flexion. The patient is instructed to flex the elbow as the examiner resists the effort. The test is considered positive if pain is elicited. (Kim et al. 2001.)

5.5 Acromioclavicular joint pathologies

Table 8. Statistical values for AC Joint tests.

Test	+LR	-LR	Reliability
Active Compression test	8.2	0.62	k = 0.22
AC Joint Palpation	1.07	0.40	n/a

Acromioclavicular joint (AC) is located between the acromion process of the scapula and the clavicle (Lippert, 2006, 96). AC joint disorders refer to any pathological state of the joint. The prevalence of these disorders in symptomatic shoulders is estimated to be 24% (Östör, Richards, Prevost, Speed & Hazleman 2005). With AC joint disorders, the Active Compression test (+LR 8.2) (also known as the O'brien Test) may be used as a confirmatory test and palpation (-LR 0.40) may be performed as a rule-out test. (Hegedus et al. 2008.) The inter-examiner reliability of the Active Compression test is poor (k = 0.22) (Cadogan, Laslett, Hing, McNair & Williams 2010).

The prevalence of AC joint disorders in symptomatic shoulders is estimated to be 24% (Östör et al. 2005). With AC joint disorders, the Active Compression test (+LR 8.2) (also known as the O'brien Test) may be used as a confirmatory test and palpation (-LR 0.40) may be performed as a rule-out test. (Hegedus et al. 2008.) The inter-examiner reliability of the Active Compression test is poor (k = 0.22) (Cadogan et al. 2010).

AC Joint Palpation is considered positive if palpation of the joint area elicits pain (Walton, Mahajan, Paxinos, Marshall, Bryant, Shier, Quinn & Murrell 2004).

In the Active Compression test the patient's arm is brought to 90 degrees of shoulder flexion, 10 degrees of horizontal adduction and full internal rotation. The patient is instructed to resist the downward force applied by the examiner. The maneuver is repeated with the patient's shoulder in full external rotation. The test is considered positive if pain is elicited in the first maneuver and reduced in the second. (Walton et al. 2004.)

6 THE MAKING PROCESS OF THE THESIS

6.1 Timeline

Figure 2. Timeline of the making process of the thesis.

The thesis process was launched in spring 2010 when we decided the subject for the thesis (figure 2). First, we started gathering background information. Once we had a basic understanding of the subject we started sketching out the written part of the thesis in September 2010. The video production began later by setting up the studio in the summer of 2011. The writing process and shooting the video were finished by October 2011. The video was completed in October 2011 by adding titles and visual elements into the cut.

6.2 Data acquisition

In the beginning of the process we familiarized ourselves with physical examination tests. Without knowledge of the number of studies that had been con-

ducted on the subject, we looked for all scientific references to physical examination tests. Starting from books and videos we narrowed down our scope, finally to individual studies. However, as we realized that our expertise and the time allocated for the thesis was insufficient for evaluating the quality of the studies and synthesizing the results, we started looking for systematic reviews on the subject. The search was done using PubMed, Google Scholar and Cochrane databases. We found six systematic reviews evaluating OSTs for the shoulder (Hegedus et al. 2008; Hughes et al. 2008; McFarland et al. 2010; Calvert et al. 2009; Munro & Healy 2009; Powell & Huijbregts 2006). However, according to The Cochrane Library only the reviews by Hegedus et al. (2008) and Hughes et al. (2008) were of sufficient quality. The first review evaluates tests for all shoulder pathologies and the latter evaluates only tests for SAIS (Centre for Reviews and Dissemination 2011a; Centre for Reviews and Dissemination 2011b.)

However, neither of these reviews is perfect. For instance in the review by Hegedus et al. (2008), the Hornblower's Sign is recommended although only one study "with small sample size and numerous design faults" had evaluated the test. Also, it ignored highly valid test combinations from the studies that were included in the review, such as the combination of Drop Arm Sign, Hawkins-Kennedy test and ERLS Test for stage III SAIS. Finally, there are some typographic mistakes in the tables which had caused the Supraspinatus/Empty Can test to be mistakenly recommended. The review by Hughes et al. (2008), on the other hand, may suffer from bias. In their work, a meta-analysis was not performed and the studies in which the tests had been shown to be ineffective were ignored in the recommendations of the review. However, as the premise of the review by Hughes et al. (2008) was to improve on the article by Hegedus et al. (2008), we took all the recommendations for SAIS tests from the article by Hughes et al. (2008). However, we excluded the Napoleon test from the Hughes et al.'s recommendations for its likeness to the Belly Press test.

Because of these defects we again considered the possibility of conducting the systematic review and meta-analysis by ourselves. In preparation for the review we studied statistical power analysis, confidence intervals and quality assess-

ment of research articles. However, once again we have to face the inadequacy of resources. What's more, there probably would have been few studies left if all the underpowered and low quality studies had been strictly excluded. Nonetheless, this experience enhanced our awareness of how to interpret studies and how future reviews could be conducted.

Due to the possibility of inaccuracy in the values reported in the reviews, we verified the data from the original studies. The instructions of how to perform the tests were taken from the same articles the reviews used for the statistical values of the tests. This way we ensured that the technique that we used is the most valid one. With tests that had been evaluated in several studies in Hughes et al.'s (2008), article, we took the statistical values from the highest quality study. If there were two or more studies with the same quality scores, we chose the statistical values from the one with the higher sample size.

6.3 The writing process

The writing process began with creating the table of contents. We wanted to sketch out the outline for the entire work before producing any actual content under the headlines. At first, we planned to include considerably more information than we finally did. In the beginning of the writing process we planned to include biomechanical basis for the tests as well as the pathophysiology behind different conditions. We abandoned the idea, however, as it was not necessary for an effective demonstration of the tests. Originally the video was designed to be educational but after understanding the time constraints of the thesis, we decided to make it instructional instead.

Before adding information under the headlines we divided the workload evenly between the two of us. We read and commented on each other's work constantly. We also revised the headlines and the content regularly. One important aspect of the workflow was a virtual space on the internet where we uploaded the latest version of the thesis every time we made changes to it. This way we always had the latest version available and the older versions in case we needed

them. Most of the writing process was done individually as we considered it more time–efficient than working together.

During the course of making the thesis, we had several meetings with the thesis workgroup and teachers to evaluate the direction of the thesis. The workgroup consisted of five peers and two physiotherapy teachers. In each of the meetings we received valuable advice and made changes accordingly. Most of the changes related to the format of the thesis. Without these meetings, the written part of the thesis would have been severely lacking in relevant information. However, the meetings did not significantly alter the actual content of the video.

Finally, as English is neither of our native languages, we used software to check the grammar and writing style of the text. Finally, we sent the thesis to our English teacher at North-Karelia University of Applied Sciences (NKUAS) to get further suggestions for correcting the text.

6.4 Video production

At first, we considered hiring a professional photographer to shoot the video. However, given our financial limitations and the relative simplicity of the task, we decided to try it ourselves. Although it was a time-consuming process, it gave us valuable skills for creating educational material in the future.

The first step in our video production was research. It encompassed acquiring facts about the subject as well as reviewing similar products that had been made in the past. This enabled us to, figuratively speaking, stand on the shoulders of other producers and try to improve their design. (LoBrutto 2002, 33–34; Rizzo 2005, 43–49.) We started looking for the strongest and weakest points of the current video products and applied these lessons in our work. In addition, we assimilated instructional material on video production to get an understanding of the rules of video making. Based on this information, we decided to try to build on the previous designs by avoiding distracting details in the background and vary the camera angles for each test. Furthermore, we decided to use sep-

arate scenes and unique camera angles for all the tests so the performing technique of the test could be seen as clearly as possible. If all the tests were to be performed consecutively as in the video by Hutchinson, the examiner might obstruct the view in some of tests. Also, we concluded that it is better to use a narration in the background than having the examiner talk while performing the tests. This was in attempt to make the video progress without unnecessary breaks and to ensure that the viewer sees the technique exactly as it should be performed in clinical practice. Finally, there seemed to be a conflict between the theory of proper studio lighting and the execution in the videos we reviewed. This was one of the aspects that we considered to hold the greatest potential for making a substantial improvement to the previous productions. (Proferes 2008, 40–51; Sawicki 2007, 137–220; LoBrutto 2002, 77–88; Brown 2007, 35–85; Box 2010, 91–108.)

Once sufficient research had been conducted, scouting was performed. Scouting signifies searching the best locations for shooting the film. (LoBrutto 2002, 33–42; Rizzo 2005, 43–49.) We had several locations to choose from, including a local physiotherapy clinic and the NKUAS. However, the problem with both of these locations was time restriction. We would have had to set up the studio and shoot the video during a weekend. Therefore, we opted to use an empty storage space that belonged to one of the authors. This way, we could set up the studio without time restrictions. The space was windowless and measured 25 m^2. The lack of external light source enabled us to have full control over the lighting of the studio. The room had enough depth but not enough width. However, we overcame this limitation with the use of a green screen, which allowed us to create the illusion of an infinite space (figure 3).

Figure 3. The studio.

A green screen was used to replace the background with a graphic design. In post–processing, the video editing software is able to remove any green color from the video clip. For instance, if an actor was wearing a green shirt, it would be made invisible. With the use of a green screen, we managed to overcome the space limitation of our studio and remove any distracting details from the background, such as shadows of the actors. To create the green screen, we used brown construction paper that we first attached to the wall with duct tape and then colored green with matte paint. A green colored paper was not used because the surface in such paper is glossy and would reflect light, making the color uneven. (Foster 2010, 139–190; Sawicki 2007, 157–198.)

For the lighting of the studio, we decided to use a basic three–point studio lighting with 800W tungsten lights and an umbrella reflector. In this setup, there are three lights around the subjects. The strongest light of these lights is called the key light. It is located behind and to the side of the camera. If this light was used by itself, only the other side of the objects in the scene would be illuminated. To make the other side visible as well, a fill light is used behind and on the opposite side of the camera in relation to the key light. It should have less intensity than the key light so that it illuminates the parts of the objects that are not visible but does not take over the key light. For the fill light, we bounced the light off an umbrella reflector in order to avoid creating distracting hard shadows on the subjects. Finally, in order to increase the definition of the objects, a light is

placed in front and to the side of the camera. This is called the rim light because it illuminates the edge of the objects, making it stand out against the background. Often, this light is colored differently to the other lights to make the scene look more interesting. In this work, we decided to use the commonly used blue tint to the light. This was achieved by using a blue semitransparent sheet of plastic in front of the rim light. At first, we wanted to use commercial quality lights borrowed from the NKUAS. However, as they would have been available only for a few days at a time we decided to look for alternatives. We bought three 400W halogen construction lights and made an umbrella reflector from a regular umbrella that we coated with aluminum foil (figure 1). To make the key light stronger, we switched on two 100W fluorescent lamps that were preinstalled in the ceiling of the room. (Jackman 2010, 109–112; Sawicki 2007, 199–220; Brown 2007, 35–85; Box 2010, 91–108.)

The video was shot in full HD quality (1920 x 1080 pixels) with Nikon 3100D digital single lens reflex (DSLR) camera. We had some previous knowledge of how to use a DSLR camera but we were not familiar with the video shooting features. We used a lot of time to find the best settings to demonstrate the tests as clearly as possible. For the camera angles, we took inspiration from the pictures in the original articles describing the tests as we considered them to be clearly portrayed. To give the viewer a better understanding of the space, we included slight camera movement in each shot. For simplicity, we shot the video holding the camera in our hands as opposed to building a track for the camera to move on. We then used a video software to reduce the shaking movement of the camera. To reinforce the perception of depth while avoiding distortion of dimensions, we used a focal length of 24–35 mm and maximum aperture size (F4.5). This way, items closer to the camera seemed bigger than items further from the camera but the difference was not exaggerated. In addition, the most important parts of the scene were in the sharpest focus and less important ones looked slightly blurry. This draws the viewer's attention to the most important elements in the scene. (Proferes 2008, 40–51; Sawicki 2007, 137–156.)

We selected the color theme for the video with a color wheel software by Adobe, called Kuler. The software includes a library of color themes based on

color theory and colors taken from images to produce harmonious combinations of colors. We selected a preset that matched our vision of the atmosphere for the video and assigned the colors to all elements of the video, such as the background, actors' clothes and text. We had to avoid, however, the color green as it could not be used anywhere else but in the green screen background (Adobe Systems Incorporated. 2011; LoBrutto 2002, 77–88; Brown 2007, 128–148.)

The video narration was recorded using a Samson C01 USB condenser microphone. This microphone was selected because of its ease of use, high quality audio recording ability and relatively low price.

Before shooting the final version of the video we made a rough edit of the entire video using video editing software. We quickly shot each scene without the green screen or other studio equipment. This helped us to get an initial idea of the scene durations as well as camera angles and settings for each test. What's more, it enabled us to preview the final product several times from the viewer's point of view. After reviewing and refining the rough edit we started shooting the final version of the video. The filming was carried out in two days. Finally, we used video software to replace the green background with a light blue background and to stabilize the camera movement. Once all the shots were processed in this manner, we cut them according to our rough edit and added title texts between the scenes.

7 DISCUSSION

Only a few instructional videos have been produced on the subject of physical examination tests for the shoulder so far (Cooke & Hegedus 2008; Hutchinson 2011). We considered there to be room for improvement in both the content of the videos and their visual presentation. Out of the videos that we are aware of, none yet demonstrate all the currently most valid tests or give information of how exactly the results of the tests should be interpreted. In a recent article that

was published in the British Journal of Sports Medicine, the Hutchinson's video series was claimed to be based on the Hegedus' systematic review, like our video (Pluim, Cingel & Kibler 2011). That is not the case, however and the claim must have originated from a misunderstanding. Furthermore, Cooke & Hegedus' (2008) videos include some of the same tests as ours but most of the recommended tests are not in it. Most likely because their videos were made before Hegedus et al.'s (2008) systematic review was finished.

As far as the visual presentation in previous videos is concerned, we think that it has been good enough for displaying the performing techniques but it could be further improved. With this work, we wanted to develop our skills in producing instructional material as well as widen our knowledge about physical examination tests. Using the green screen or 3D technology may not have been necessary to show the performing techniques but these matters helped in taking our work beyond the work of others. Perhaps some of it may later be found to be unnecessary or confusing but it will serve as grounds for others to build on.

In retrospect, we are satisfied with both the content and the visual presentation of the video. We found that producing a fairly high-quality video doesn't necessarily require a big budget. We were able to keep a high work ethic throughout the process and avoid taking shortcuts although making compromises would have been easier at times. An instance of these times was the decision to not only go with Hegedus et al.'s (2008) recommendations for the tests but to look for all research on the subject. In the process, we learned a great deal about finding reliable information, evaluating research and clinical tests and understanding the role of OSTs in the examination process.

There is a clear trend that the higher the quality of the study the lower the validity of the test is. This can also be observed when calculating sufficient sample size for a study–the lower the validity value that is being looked for, the higher the sample size (Carley, Dosman, Jones & Harrison 2005). This raises a suspicion that the current research may not give an accurate estimation of the validity of the tests. Unfortunately, as with other research it has to be concluded that more research is warranted to make more certain recommendations.

Another common trend is the inconsistency of the validity and reliability of the tests. Often the originators of the test get excellent validity scores for the tests in their own research but when the research is conducted by others, the validity is substantially lower. (Hegedus et al. 2008; May et al. 2010.) This may be the result of methodological defects in the originators' study or possibly a sign of low reliability in case the others did not repeat the same test maneuver. In the original articles the written descriptions of the tests were brief and lacking detail. In our opinion, this could be one reason for the low reliability of the tests and lower validity values in studies conducted after the original study. Anyway, the overall usefulness of the tests may be questioned.

From an ethical point of view, one may question if it is acceptable to make a video of clinical tests that have not been convincingly proven to be valid and reliable. Such a product may promote making of incorrect hypotheses and thus incorrect treatment choices. This could have damaging consequences for patients. However, we took every step to ensure that all of our decisions about the video's content were ones that take physiotherapists' closer to making more accurate hypotheses of the origin of the patient's pain. All the information is based on peer–reviewed reviews and studies. We attempted to pass the information as unchanged as possible while, of course, avoiding copying any author without giving them due credit. The performing instructions of the tests are quite similar to those used in the original articles but because of the nature of the subject they cannot be changed too much. Another ethical consideration of our thesis was free-riding since this thesis was made by two contributors. Nevertheless, we were able to divide the workload equally between the two of us and the contribution of both authors was of similar standard.

We noticed inconsistencies in not only the performing techniques in the studies evaluating physical examination tests but also in the videos. In Hutchinson's (2011) video, some of the tests are performed in a slightly different manner than how they were performed in the highest quality studies. For instance, the Hawkins-Kennedy test is performed so that the patient's arm is in 90 degrees of shoulder flexion and about 45 degrees of horizontal abduction which differs

from the technique that we recommend. It is also stated that pain in the Biceps load II is a sign of biceps tear while it should be used to detect SLAP tears. Finally, the Apprehension test performing technique resembles a Posterior Clunk test because the humeral head is pressed anteriorly with the thumb as the patient's arm is brought into the apprehension position. Again, this way of performing the test is different from the one that has been researched and recommended. Overall, it seems that some of the recommendations may be based on Dr. Hutchinson's clinical experience. Many of the demonstrated tests could be valid but without evidence, recommending them may not be warranted.

The QUADAS score was used in Hegedus et al.'s (2008) to assess the methodological quality of the studies. QUADAS consists of 14 questions and each is awarded with a point. Hegedus et al. (2008) considered studies with 10 points or more to be of sufficient quality. However, in our opinion some methodological issues are more important than others. For instance, if a study is methodologically sound in every other way, but the study population was not representative of the population that will receive the test in clinical setting, the validity scores are greatly misleading. To overcome this problem, only QUADAS items that are relevant to the study should be included and a rating method by total score should not be used (Whiting, Rutjes, Reitsma, Bossuyt & Kleijnen 2003.) Some high quality studies may get low QUADAS scores only because the articles failed to describe clearly the methods they used. Therefore, the QUADAS may not evaluate only the quality of the study but also the quality of the study report.

In the future, more studies with high methodological quality and sufficient sample size should be conducted to evaluate the validity and reliability of the most common OST's for the shoulder. This could be accomplished by referring to the QUADAS score in the design process and calculating the required sample size with the use of the nomogram by Carley et al. (2005). Also, consistent performing techniques, such as those presented in our work, should be used in the studies.

OST's hold potential for widely available, practical and inexpensive way of making a diagnosis. However, due to the low amount of high quality studies availa-

ble, their use is not reliable for making diagnoses yet. On the other hand, it appears that the higher quality the study, the lower the validity score of the test is. However, even if future high quality studies found the validity to be low for all OSTs, the greater number of tests could make a reliable diagnosis possible with the use of a nomogram.

REFERENCES

Adobe Systems Incorporated. 2011. Adobe Kuler. http://www.adobe.com/products/kuler. 5.10.2011.

Agoritsas, T., Courvoisier, D.S., Combescure, C., Deom, M., Perneger, T.V. 2010. Does prevalence matter to physicians in estimating post-test probability of disease? A randomized trial. Journal of General Internal Medicine. 26 (4), 373–378.

Attia, J.R., Nair, B.R., Sibbritt, D.W., Ewald, B.D., Paget, N.S., Wellard, R.F., Patterson, L., Heller, R.F. 2004. Generating pre-test probabilities: a neglected area in clinical decision making. The Medical Journal of Australia. 180 (9), 449–454.

Auge, W.K., Fischer, R.A. 1998. Arthroscopic distal clavicle resection for isolated atraumatic osteolysis in weight lifters. The American Journal of Sports Medicine. 26 (2), 189–192.

Barth, H.R.J., Burkhart, S.S. & De Beer, F.J. 2006. The bear–hug test: A new and sensitive test for diagnosing a subscapularis tear. The Journal of Arthroscopic and Related Surgery. 22 (10), 1076–1084.

Bongers, P.M. 2001. The cost of shoulder pain at work. British Medical Journal. 322 (7278), 64–65.

Box, H. 2010. Set Lighting Technician's Handbook. UK: Focal Press.

Brown, B. 2007. Motion Picture and Video Lighting. Oxford, UK: Focal Press.

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. 16 (2), 131–135.

Calvert, E., Chambers, G.K.., Regan, W., Hawkins, R.H., Leith, J.M. 2009. Special physical examination tests for superior labrum anterior posterior shoulder tears are clinically limited and invalid: a diagnostic systematic review. Journal of Clinical Epidemiology. 62 (5), 558–563.

Campbell, M.J., Machin, D. & Walters, S.J. 2007. Medical Statistics: A Textbook for the Health Sciences. Chichester: Wiley–Interscience.

Carley, S., Dosman, S., Jones, S.R., Harrison, M. 2005. Simple nomograms to calculate sample size in diagnostic studies. Emergency Medicine Journal. 22 (3), 180–181.

Centre for Reviews and Dissemination. 2011a. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests (Structured abstract). Cochrane Database of Abstracts of Reviews of Effects. (3).

Centre for Reviews and Dissemination. 2011b. Most clinical tests cannot accurately diagnose rotator cuff pathology (Structured abstract). Cochrane Database of Abstracts of Reviews of Effects. (4).

Cooke, C. & Hegedus, E. 2008. Orthopaedic physical examination tests–an evidence based approach. Pearson Education Inc. Video CD.

Davidson, M. 2002. The interpretation of diagnostic test: a primer for physiotherapists. Australian Journal of Physiotherapy. 48 (3), 227–232.

Deeks J.J. & Altman D.G. 2004. Diagnostic tests 4: likelihood ratios. British Medical Journal. 329 (7458), 168–169.

Fagan, T.J. 1975. Letter: Nomogram for Bayes Theorem. The New England Journal of Medicine. 293 (5), 257.

Farber, A., Castillo, L., Clough, M., Bahk, M., McFarland, E. 2006. Clinical assessment of three common tests for traumatic anterior shoulder instability. The Journal of Bone and Joint Surgery. 88 (7), 1467–1474.

Fletcher, R.H. & Fletcher, S.W. 2005. Clinical Epidemiology: The Essentials. Versailles: Lippincott Williams & Wilkins.

Foster, J. 2010. The Green Screen Handbook: Real–World Production Techniques. Indiana, US: Wiley Publishing.

Fritz, J.M. & Wainner, R.S. 2001. Examining diagnostic tests: an evidence–based perspective. Physical Therapy. 81 (9), 1546–1564.

Gadotti I, Vieira E, Magee D. 2006. Importance and Clarification of Measurements Properties in Rehabilitation. Revista Brasileira de Fisioterapia. 10 (2), 137–146.

Habermeyer, P., Magosch, P. & Lichtenberg, S. 2006. Classifications and scores of the shoulder. Heidelberg: Springer Berlin.

Hegedus, E.J., Goode, A. Campbell, S., Morin, A., Tamaddoni, M., Moorman C.T. & Cook, C. 2008. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. British Journal of Sports Medicine. 42 (2), 80–92.

Hughes, C.P., Taylor, F.N. & Green, A.R. 2008. Most clinical tests cannot accurately diagnose rotator cuff pathology: a systematic review. Australian Journal of Physiotherapy 54 (3), 159–170.

Hutchinson, M. 2011. Shoulder Exam. http://www.youtube.com/user/BJSMVideos. 28.8.2011.

Jackman, J. 2010. Lighting for Digital Video and Television. Oxford, UK: Focal Press.

Khan, K.S., Chien, P.F. 2001. Evaluation of a clinical test. I: assessment of reli–ability. BJOG: An International Journal of Obstetrics and Gynaecology. 108 (6), 562–567.

Kim, S., Ha, K., Ahn, J., Kim, S. & Choi, H. 2001. Biceps Load Test II: A clinical test for SLAP lesions of the shoulder. Arthroscopy. 17 (2), 160–164.

Kim, T., Queale, W., Cosgarea, A. & McFarland, E. 2003. Clinical features of the different types of SLAP lesions: an analysis of one hundred and thirty–nine cases. Journal of bone and joint surgery. 85 (1), 66–71.

Kuhn, J.E. 2010. A new classification system for shoulder instability. British Journal of Sports Medicine 44 (5), 341–346.

Lewisa, A., Kitamurab, T. & Bayley J.I.L. 2004. (ii) The classification of shoulder instability: new light through old windows! Current Orthopaedics. 18 (2), 97–108.

Lippert, L.S. 2006. Clinical kinesiology and anatomy. 4th edition. Philadelphia: F.A. Davis company.

Lo, I., Nonweiler, B., Woolfrey, M., Litchfield, R. & Kirkley, A. 2004. An evaluation of the Apprehension, Relocation, and Surprise Tests for anterior shoulder Instability. American journal of sports medicine. 32 (2), 301–307.

LoBrutto, V. 2002. The Filmmaker's Guide to Production Design. New York: Allworth Press.

Magee, J.D. 2006. Orthopedic Physical Assessment, 4th edition. Canada: Elsevier.

Maritz, N,G,J. & Oosthuizen, P.J. 2002. Diagnostic criteria for acromioclavicular joint pathology. Journal of Bone and Joint Surgery. 84B (Suppl.1), 78.

Marx, R.G., Menezes, A., Horovitz, L., Jones, E.C., Warren, R.F. 2003. A comparison of two time intervals for test–retest reliability of health status instruments. Journal of Clinical Epidemiology. 56 (8), 730–735.

May, S., Chance–Larsen, K., Littlewood, C., Lomas, D. & Saad M. 2010. Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review. Physiotherapy. 96 (3), 179–190.

McFarland, E.G., Garzon–Muvdi, J., Jia, X., Desai, P., Petersen, S.A. 2010. Clinical and diagnostic tests for shoulder disorders: a critical review. British Journal of Sports Medicine. 44 (5), 328–332.

McGee, S. 2002. Simplifying likelihood ratios. Journal of General Internal Medicine. 17 (8),646–649.

Michener, L.A., Walsworth, M.K. & Burnet, E.N. 2004. Effectiveness of rehabilitation for patients with subacromial impingement syndrome: a systematic review. Journal of Hand Therapy. 17 (2), 152–164.

Moen, M.H., Jan de Vos, R., Ellenbecker, T.S. & Weir, A. 2010. Clinical tests in shoulder examination: how to perform them. British Journal of Sports Medicine. 44 (5), 370–375.

Moosmayer, S., Smith, H.J., Tariq, R. & Larmo, A. 2009. Prevalence and characteristics of asymptomatic tears of the rotator cuff: an ultrasonographic and clinical study. The journal of bone and joint surgery, British Volume. 91 (2), 196–200.

Munro, W., Healy, R. 2009. The validity and accuracy of clinical tests used to detect labral pathology of the shoulder—a systematic review. 14 (2), 119–130.

Neer, CS. 1983. Impingement lesions. Clinical Orthopaedics and Related Research. 173, 70–77.

Ozaki, J., Fujimoto, S., Nakagawa, Y., Masuhara, K. & Tamai, S. 1988. Tears of the rotator cuff of the shoulder associated with pathological changes in the acromion. A study in cadavera. The Journal of Bone and Joint Surgery, American volyme. 70 (8), 1224–1230.

Park, H.B., Yokota, A., Gill, H.S., El Rassi, G. & McFarland, E.G. 2005. Diagnostic Accuracy of Clinical Tests for the Different Degrees of Subacromial Impingement Syndrome. The Journal of Bone and Joint Surgery. 87 (7), 1446–1455.

Parsons, S., Breen, A., Foster, N.E., Letley, L., Pincus, T., Vogel, S., Underwood, M. 2007. Prevalence and comparative troublesomeness by age of musculoskeletal pain in different body locations. Family Practice. 24 (4), 308–316.

Pluim, B.M., van Cingel, R.E.H. & Kibler, W.B. 2011. Shoulder to shoulder: stabilizing instability, re–establishing rhythm, and rescuing the rotators! British Journal of Sports Medicine, 44 (5), 299.

Porta, M. 2008. A Dictionary of Epidemiology, 5th edition. New York: Oxford University Press.

Portney L.G. & Watkins M.P. 2000. Foundations of Clinical Research: applications to practice, 2nd edition. New Jersey: Prentice Hall Health. 560–567.

Powell, J.W. & Huijbregts, P.A. 2006. Concurrent Criterion–Related Validity of Acromioclavicular Joint Physical Examination Tests: A Systematic Review. The Journal of Manual & Manipulative Therapy. 14 (2), E19–E29.

Proferes, N. 2008. Film Directing Fundamentals, Third Edition: See Your Film Before Shooting. Oxford, UK: Focal Press.

Reilly, P., Macleod, I., Macfarlane, R., Windley, J. & Emery, R.J.H. 2006. Dead men and radiologists don't lie: a review of cadaveric and radiological studies of rotator cuff tear prevalence. Annals of the royal college of surgeons of England. 88 (2), 116–121.

Riddle D. & Stratford P.W. 1999. Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. Physical Therapy. 79 (10), 939–948.

Rizzo, M. 2005. The Art Direction Handbook for Film. Oxford, UK: Focal Press.

Rothstein, J.M. 2001. Sick and Tired of Reliability? Physical Therapy. 81 (2), 774–775.

Sawicki, M. 2007. Filming the Fantastic: A Guide to Visual Effects Cinematography. Oxford, UK: Focal Press.

Shrout, P.E., Fleiss, J.L. 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 86 (2), 420–428.

Silman, A.J. & Macfarlane, G.J. 2002. Epidemiological Studies: A Practical Guide. Cambridge: Cambridge University Press.

Sim, J. & Wright C.C. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy. 85 (3), 257–268.

Simon, S. 2002. Likelihood ratio slide rule. Children's Mercy Hospitals and Clinics. http://www.childrensmercy.org/stats/sliderule.asp. 30.8.2011.

Sizer, P.S., Phelps, V., Gilbert, K. 2003. Diagnosis and management of the painful shoulder. Part 2: examination, interpretation, and management. Pain Practice. 3 (2), 152–185.

Snyder, S., Karzel, R., Del Pizzo, W., Ferkel, R. & Friedman, M. 1990. SLAP le–sions of the shoulder. Arthroscopy. 6 (4), 274–279.

Stemler, S.E. 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9 (4). http://PAREonline.net/getvn.asp?v=9&n=4. 8.4.2011.

Tempelhof, S., Rupp, S. & Seil, R. 1999. Age–related prevalence of rotator cuff tears in asymptomatic shoulders. Journal of Shoulder and Elbow Surgery. 8 (4), 296–299.

Tennent, T.D., Beach, W.R., Meyers, J.F. 2003. A review of the special tests associated with shoulder examination. Part I: the rotator cuff tests. American Journal of Sports Medicine. 31 (2), 154–160.

Van der El A. 2010. Orthopaedic Manual Therapy Diagnosis: Spine and Temporomandibular Joints. Sudbury, MA: Jones & Bartlett Publishers.

Viera, A.J., Garrett, J.M. 2005. Understanding interobserver agreement: the kappa statistic. Family Medicine. 37 (5), 360–363.

Wainner. R.S. 2003. Reliability of the clinical examination: how close is "close enough"? Journal of Orthopaedic & Sports Physical Therapy. 33 (9), 488–491.

Walton, J., Mahajan, S., Paxinos, A., Marshall, J., Bryant, C., Shier, R., Quinn, R. & Murrell, G. 2004. Diagnostic values of tests for acromioclavicu-

lar joint pain. The Journal of Bone and Joint Surgery. 86 (4), 807–812.

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.

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 included in systematic reviews. BioMedCentral. Medical research methodology. 3 (25).

Wolf, M. & Agrawal, V. 2001. Journal of Shoulder and Elbow Surgery. 10 (5), 470–473.

Worland, R., Lee, D., Orozco, C., SozaRex, F. & Keenan, J. 2003. Correlation of age, acromial morphology, and rotator cuff tear pathology diagnosed by ultrasound in asymptomatic patients. Journal of Southern Orthopaedic Association. 12 (1), 23–26.

World Health Organization. 2011. International Classification of Functioning, Disability and Health. Online version. http://apps.who.int/classifications/icfbrowser/. 28.8.2011.