

Muscle strength in adults with spinal cord injury: A systematic review of manual muscle testing, isokinetic and hand held dynamometry clinimetrics.

Master of Clinical Science
February 2014
Kerry Peek BSc (Hons), PG Cert.
Student number a1642099
The Joanna Briggs Institute,
School of Translational Health Science,
Faculty of Health Sciences,
The University of Adelaide,
Level 1, 115 Grenfell Street,
SA. 5005.
kerry.peek@adelaide.edu.au

Secondary reviewer:**Prapaphan Pensuk RN, B.N.S., M.N.S**

The Joanna Briggs Institute,
School of Translational Health Science,
Faculty of Health Sciences,
The University of Adelaide,
SA. 5005.

prapaphan.pensuk@adelaide.edu.au

Supervisors:**Dr Suzanne Robertson-Malt, BHSc (Hons), PhD**

The Joanna Briggs Institute,
School of Translational Health Science,
Faculty of Health Sciences,
The University of Adelaide,
SA. 5005.

suzanne.robertson-malt@adelaide.edu.au

Dr Zachary Munn, PhD, GradDipHlthSci, BMedRad

The Joanna Briggs Institute,
School of Translational Health Science,
Faculty of Health Sciences,
The University of Adelaide,
SA. 5005.

zachary.@adelaide.edu.au

Additional support:**Dr Catalin Tufanaru, MD, MPH**

The Joanna Briggs Institute,
School of Translational Health Science,
Faculty of Health Sciences,
The University of Adelaide,
SA. 5005.

catalin.tufanaru@adelaide.edu.au

Table of Contents

Title page: Muscle strength in adults with spinal cord injury: A systematic review of manual muscle testing, isokinetic and hand held dynamometry clinimetrics	1
Supervisors	2
Contents	3
Abstract	5
Student Declaration Statement	6
Acknowledgements	7
Chapter 1: Introduction	8
Context of the review	8
Spinal Cord Injury	8
Muscle strength	9
Muscle strength assessment	10
Methods for conducting muscle strength assessments	11
Clinimetrics	12
Statement of review question	16
Overview of the science of evidence synthesis	17
Methodological issues, critical appraisal, data extraction and data synthesis	18
Definition of terms	20
Chapter 2: Methods	22
Review objective	22
Criteria for considering studies for this review	23
<i>Types of studies</i>	23
<i>Types of participants</i>	23
<i>Focus of the review</i>	23
<i>Outcome measures</i>	23
Review methods	24
<i>Search strategy</i>	24
<i>Assessment of methodological quality</i>	24
<i>Data extraction</i>	25
<i>Data synthesis</i>	25
Chapter 3: Results	26
Description of studies	26
Review findings/ results	29
Reliability	33
<i>Kappa</i>	33
<i>Intraclass correlation coefficient</i>	34
<i>Confidence intervals</i>	35
<i>Manual muscle testing</i>	35
<i>Isokinetic dynamometry</i>	38

<i>Hand held dynamometry</i>	40
Validity	44
<i>Pearson product moment correlation coefficient</i>	44
<i>Spearman rank correlation coefficient</i>	44
<i>t-test</i>	44
<i>p-value</i>	44
Responsiveness	49
Interpretability	50
Key findings	52
Hand held dynamometry	55
Chapter 4: Discussion	58
Methodological quality	58
Clinimetrics	58
Reliability	59
Validity and Responsiveness	61
Interpretability	63
Additional Variables	64
Hand held dynamometry	66
Assumptions	69
Limitations	69
Delimitations	69
Chapter 5: Conclusions	71
Implications for practice	71
Implications for research	72
General conclusion	73
References	76
Appendix I Data Extraction	79
Appendix II COSMIN tool for methodological quality	80
Appendix III Search Strategy	96
Appendix IV Included studies	97
Appendix V Excluded studies	120

Abstract

Objectives

The objectives of this systematic review were to synthesise the best available evidence regarding the clinimetrics for manual muscle testing (MMT), isokinetic dynamometry (ID) and hand held dynamometry (HHD) in the assessment of muscle strength in adults with spinal cord injury (SCI) and determine whether there is research evidence to recommend HHD as the standard tool for measuring muscle strength in adults with SCI.

Inclusion Criteria:

Only studies related to adults with SCI and MMT and/or ID and/or HHD were included.

Search Strategy

The search sought to identify any relevant English language published or unpublished studies via a three step search strategy.

Methodological quality

Two independent reviewers assessed the methodological quality of the studies using the quality evaluation tool consensus-based standards for the selection of health status instruments (COSMIN).

Data collection

An original data extraction form was developed to extract quantitative data from the included studies.

Data synthesis

It was not appropriate to conduct a meta-analysis due to the heterogeneity of the included studies. Therefore, the results are presented in narrative text including raw data as presented in the included studies as well as the contextual data.

Results

Eleven studies met the inclusion criteria of this systematic review. The results demonstrated that MMT showed varying inter-tester reliability over 10 muscle groups tested, ID demonstrated good reliability for the shoulder but not the elbow, HHD showed good reliability and validity for the upper limb and trunk, as well as good results for responsiveness and interpretability. Positive correlations were seen between MMT, ID and HHD particularly at the lower MMT grades. However, change in muscle strength scores seen on ID and HHD testing were not always correlated with changes in MMT grade. Significant overlapping of scores was seen between MMT and HHD particularly for grades 4 and 5 with MMT unable to detect subtle changes in muscle strength compared with dynamometry.

Conclusions

In conclusion, when considering the clinimetrics of the 3 methods for assessing muscle strength in adults with SCI there does appear to be support in the literature to recommend the wider application of HHD compared with MMT and ID.

Student Declaration Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give my permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed:

Kerry Peek

Dated:

Acknowledgements

I would like to thank my supervisors, Dr Robertson-Malt and Dr Munn, for their support and advice during my Master's candidature. A special thank you to Dr Catalin Tufunaru for his invaluable assistance.

My secondary reviewer, Prapaphan Pensuk, for her assistance with methodological quality assessment, reassurance and time.

My family, Brendan, Harry and Charlie, who encouraged me and gave me time and support to get on with it!

Chapter 1: Introduction

Context of the review

The assessment of muscle strength is a routine part of any thorough objective examination carried out by a physiotherapist worldwide, and in particular following spinal cord injury (SCI). There are three main methods for determining muscle strength, manual muscle testing (MMT), isokinetic dynamometry (ID) and hand held dynamometry (HHD). Technology has seen enormous change to the everyday lives of individuals over the last 40 years. Health professionals and physiotherapists alike are not immune to these developments; numerous sophisticated technological devices have been successfully incorporated into clinical practice. This thesis aims to systematically review the best available evidence on the clinimetrics of three methods for assessing muscle strength, MMT, ID and HHD in adults with SCI and to determine whether there is support in the research to recommend the use of HHD as the standard measurement tool. My interest in this area is two fold. I grew up close to the National Spinal Injuries Centre at Stoke Mandeville Hospital, UK (in fact I ran a marathon to raise money for the upgrade of the centre in 2003) and undertook elective physiotherapy placement at the Spinal Cord Injury Centre at the Royal National Orthopaedic Hospital in Stanmore, UK which incorporated the Spinal Injury Games at Stoke Mandeville. Secondly, I have worked extensively with dynamometry, both ID and HHD whilst in the UK in the assessment and treatment of spinal and other musculo-skeletal injuries, where the use of quantifiable data became inherent to clinical practice and patient outcome measures. The decision to specify this review to SCI rather than any other musculo-skeletal injury or even normal participants, was based on the assumption that SCI has the potential to demonstrate muscle functioning throughout the entire range. Muscles above the injury are usually unaffected and therefore can be considered 'normal' whereas muscles below a complete SCI can be considered absent in function, giving the two ends of the spectrum. However, with incomplete SCI, the function of any particular muscle or muscle group can sit anywhere between. Given this range, it is hoped that the results might be discussed and comparisons made, at least in part, to other research and systematic reviews in this field using differing conditions and therefore, encourage a wider application to clinical practice.

This first chapter aims to provide the background and justify the importance of this review, establish the historical context and the development of quantifiable assessment of muscle strength related to SCI.

Spinal Cord Injury

SCI is often defined as damage either through trauma or disease to the spinal cord which results in a loss of function such as motor control and sensation. If the injury to the spinal cord occurs in the cervical spine the resultant loss of function will involve the upper limbs, trunk and lower limbs, which is termed tetraplegia. The term paraplegia, relates to functional loss which involves the trunk and lower limbs only. The American Spinal Injury Association (ASIA) further defines SCI as 'complete' or 'incomplete' using the ASIA Impairment Scale.¹ According to this scale a 'complete' SCI presents with no preservation of motor or sensory function below the sacral segments of S4-5 (in other words no rectal sensation or voluntary anal sphincter control). Whereas an 'incomplete'

SCI has preserved sensory function below the spinal level of injury including S4-5, with or without motor function. However, there are 3 categories of ‘incomplete’ lesions depending on how much motor or sensory preservation is seen below the neurological level of injury, with the severity of impairment being more pronounced in ASIA grade B compared to ASIA grade D. (see table 1 below).¹ In Australia, the SCI prevalence rate among individuals 15 years and older has been reported as 681 per million population.² This rate is considered lower than the United States but higher than a number of European countries. With increasing survival and little variation in the incident case number evident from data collected since 1986, the prevalent population of spinal cord injuries in Australia is increasing.²

SCI is arguably one of the most life altering medical conditions; the impact of which can have consequences in all facets of function and requires extensive rehabilitation.

Table 1: ASIA Impairment Scale¹

A = Complete:	No motor or sensory function is preserved in the sacral segments S4-S5.
B = Incomplete:	Sensory but not motor function is preserved below the neurological level and includes the sacral segments S4-S5.
C = Incomplete:	Motor function is preserved below the neurological level, and more than half of key muscles below the neurological level have a muscle grade less than 3.
D = Incomplete:	Motor function is preserved below the neurological level, and at least half of key muscles below the neurological level have a muscle grade of 3 or more.
E = Normal:	motor and sensory function are normal

Muscle Strength

Muscle strength can be defined as the maximal force generated by a muscle at a specific velocity.³

More specifically a muscle contraction can be divided into a number of different types:

Isotonic contraction: Muscle tone remains the same throughout the range of movement and should include both concentric and eccentric muscle contractions.³

Concentric contraction: The origin and insertion of the muscle’s attachments move closer together (i.e. shortening of muscle fibres) during the contraction.³

Eccentric contraction:	The origin and insertion of the muscle's attachments move further apart (i.e. lengthening of the muscle fibres) during the contraction. ³
Isometric contraction:	The joint angle and muscle length does not alter during the muscle contraction. ³
Isokinetic contraction:	This is a dynamic action where equal speed and resistance is provided by a specialist machine throughout the range of movement as the muscle length changes. ³

Isometric contractions have been reported as perhaps the safest form of muscle contraction and therefore, pose the lowest risk of injury but perhaps a smaller potential for strength gain.³ Whereas eccentric actions require increased control of the lengthening action of the muscle which results in greater muscle tension, an increased potential for injury but an increased potential for muscle hypertrophy.³

Muscle Strength Assessment

Physiotherapists and many other health professionals routinely conduct muscle strength tests of their patients as part of the objective examination. Following SCI, the determination of muscle strength is important for a number of reasons including neurological classification of injury, therapeutic planning and outcome evaluation.⁴

Accurate baseline assessment of muscle strength in the initial examination of an adult with SCI is important for physiotherapists as it will direct them in their treatment plan and treatment goal setting process. SCI and their subsequent impairment tend to vary between individuals particularly with the number of incomplete injuries. With an incomplete injury the individual may have varying muscle activity below the main spinal level of injury which needs to be carefully assessed and measured so that no muscle activity is missed which could have a substantial impact on the individual's treatment and exercise focus and therefore, functional recovery. Accurate muscle strength testing is necessary to highlight areas of muscle weakness or indeed strength as well as muscle imbalances between agonistic and antagonistic muscle groups in order to direct the exercise component of the rehabilitation program.

Accurate retesting of muscle strength is also required at follow up appointments as a indicator of the effect of physiotherapeutic interventions. The progressive evaluation of physiotherapy treatment outcomes is an integral part of professional accountability and is a requirement of the Australian Physiotherapy Competency Standards.⁵ There is an emphasis on clinical outcome measures for physiotherapists during rehabilitation of their patients and accurate measurement of muscle strength can provide quantifiable data in support of patient outcomes. A significant percentage of traumatic spinal cord injuries are caused through road traffic accidents. Funding for physiotherapy through external agencies such as the Motor Accident Commission require the physiotherapist to provide evidence of improvement and state the outcome measures used in support of the need for on-going physiotherapy. Muscle strength data can assist with this.

Methods for conducting muscle strength assessment

A number of methods are available for assessing muscle strength. MMT commonly takes the form of requesting the patient to perform a muscle activity which is then graded by its ability to act in a gravity eliminated position, against gravity or the resistance applied by the therapist.⁴ The therapist may then record this subjectively utilizing their experience as to whether they felt a muscle or action was impaired or not. They may quantify this muscle activity using the Scale developed by the Medical Research Council (MRC), which uses an ordinal scale of 0-5 (see table 2). However, this scale also relies somewhat on the therapist's experience. In addition, the use of the MRC scale maybe considered valid within an individual but not across individuals. The scale is nominal in design and therefore the difference between two points on the scale i.e. level 2 and 4 is not quantifiable. For example level 4 is not equal to twice as much strength as the level 2. The reliability of MMT is increased by the use of standardised testing positions for each muscle particularly when classifying SCI using the International Standards for Classification of Spinal Cord Injury (revised in 2000).⁶

The table below gives the MRC scale for assessing muscle strength.

Table 2: Medical Research Council Scale⁷

Grade	MRC grade of muscle strength
0	No movement
1	Flicker of movement only
2	Movement possible when assisted by gravity or gravity is eliminated
3	Movement possible against gravity but without imposed resistance
4	Movement possible against gravity and some resistance
5	Normal

As early as 1916 the dissatisfaction with imprecise and subjective measures of muscle strength was reported.⁸ By 1956, evidence for the use of HHD also known as myometers to objectify muscle strength tests was being demonstrated.⁸ HHD were promoted as portable, inexpensive devices that can be used to obtain more objective measures of muscle strength.⁹ The portability of HHD enables testing of multiple muscles, performed at several points in the range of movement, over a short period and allows for bedside testing.¹⁰

Further developments in the use of dynamometry over the last 35 years or so has seen the emergence of large computerised testing devices known as ID, which was reported as the 'gold standard' of measuring muscle strength in a systematic review by Stark et al. in 2011.¹¹ The main differences between HHD and ID are related to ease of use, cost (ID can cost 40 times more than a HHD), size and access.

MMT in contrast requires no equipment and therefore is the most widely used method clinically. MMT can also be performed quickly, easily understood and utilised even

when severe muscle weakness is present.¹⁰ However, where MMT uses ordinal data, dynamometry uses the highest level of measurement, the ratio scale, because it uses a scale where the units of measure have equal intervals with a true zero, therefore, the data gathered between and across individuals allows for more mathematical and statistical comparisons to be drawn.

MMT is graded by the muscle's ability to act with or without gravity in an isotonic contraction. The examiner may ask the patient to perform an isometric contraction where resistance is applied to discern between grades 4 and 5. HHD measures muscle strength via an isometric contraction although more recent advanced machines may also be able to measure isotonic contractions through a range of movement. On the other hand ID is able to test muscle strength during all forms of muscle contractions, isometric, isotonic and isokinetic.

When looking specifically at isometric contractions, particularly with reference to HHD, there are two types of testing techniques; 'make' or 'break'. The 'make' technique requires the patient to exert a maximal contraction against the HHD, resisted by the examiner in a fixed position.¹² Whereas the 'break' technique requires the examiner to overpower a maximal effort by the patient, thereby producing a measurement of eccentric muscle strength.¹² Eccentric contractions and therefore, the 'break' technique generally produce higher strength values.¹²

The type of muscle contraction the examiner wishes to measure is dependent on the patient's requirements, whether looking at baseline strength (where isometric muscle contractions maybe the most appropriate) or function (utilising isotonic muscle contractions) or sophisticated muscle ratios (using isokinetic testing). However, the different types of measurement should not be interchanged when comparing test and re-test data. With regard to this review the types of HHD, make or break technique will be stated just as the ID setting for greater clarity when comparing results, for example, whether the ID used isometric or isokinetic muscle contractions and the speed of movement observed.

The other area which has not been discussed is grip strength dynamometers. It is not the purpose of this review to include grip strength as this falls within a very specialised area. Mafi et al.¹³ conducted a systematic review of hand dynamometry and it's role in hand trauma assessment which found acceptable to high levels of reliability. Grip strength measurement instruments are more specific than HHD. They can be hydraulic, pneumatic, mechanical or strain gauge types and are generally only used to measure grip strength and not any other joint in the body, which is the reason for exclusion.

Clinimetrics

Clinimetrics is a methodological discipline that focuses on the quality of measurement in medical and health care research. The quality of measurement includes both the quality of the measurement instrument and the quality of performing the actual measurements.¹⁴

The objective of this systematic review is to consider the clinimetrics for three methods of assessing muscle strength specific to SCI. The methodological quality of the studies to be included will be based on the Consensus-based Standards for the selection of health

Measurement INstruments (COSMIN) checklist. The development of the COSMIN tool arose from an international initiative to improve the quality assessment and selection of health measurement instruments.¹⁴ This initiative resulted in the development and design of the COSMIN critical appraisal tool (checklist). The checklist was developed in an international Delphi study as a multidisciplinary, international collaboration.¹⁴ For the purpose of this review a number of terms have been defined in accordance with COSMIN as these will be the clinimetric domains:

Reliability of the measurement of muscle strength: The degree to which the measurement of muscle strength is free from measurement error. This includes repeated measurement under several conditions such as over time (test-retest), by different people (inter-tester) or by the same person (intra-tester). For the purposes of this review reliability will be considered for MMT, ID and HHD.¹⁴

Validity of the muscle strength test: The degree to which the muscle strength test measures what it purports to measure. This includes face validity (the degree to which the muscle strength test adequately reflects a measurement of muscle strength) and criterion validity (the degree to which the measurements data is an adequate reflection of a 'gold standard').¹⁴

Responsiveness of the muscle strength test: The ability of the muscle strength test to detect muscle strength changes over time.¹⁴

Interpretability of the muscle strength test: The degree to which one can assign clinical meaning to the quantitative scores or change in scores in muscle strength.¹⁴

The role of measurement in physiotherapy practice cannot be overstated.⁵ Physiotherapists are moving towards choosing assessment procedures that have been shown in research to be reliable and valid. Such instruments increase the likelihood of being able to measure accurately the impact of an intervention or change over time.⁵

Clinically it is important to understand the clinimetrics of the method being used to assess an individual's muscle strength. For example, a literature review by Sisto and Dyson-Hudson¹⁰ reported that instrumented strength measures produce greater reliability than MMT, generally because of their measurement quality. However, in the case of ID which may not be available to the clinician, either due to access or the time taken to set up or calibrate the machine, it may be more practical to use MMT or HHD.

When considering an instrument's validity it is important to question whether 'we are measuring what we think we are measuring'. When further exploration is made into the relationship between validity and reliability, it is important to understand the main differences between the two terms. In the example of measuring muscle strength, the method used may be considered in the research to be reliable in that the same result was measured three times on the same muscle of the same patient. However, the measurement may not be valid. It is possible that although the measurement was the same each time, and therefore reliable, the measurement may not accurately reflect the true muscle strength score i.e. if the dynamometer was not calibrated correctly or the MMT overestimated or underestimated the true muscle strength.¹⁵ It has been reported that as the MRC scale is non-linear and relatively insensitive to change, that a grade 4 strength may be assigned with as little as 10% of predicted normal strength.¹² This would indicate

that MMT may be reliable but not necessarily valid. Therefore, reliability does not presuppose validity, in addition, if an instrument yields unreliable measurements, this will set limits to its validity.¹⁶ The concept of error is another area to consider. If the muscle strength measurements are closely grouped but not actually reflecting the true result (such as in the case of a non-calibrated dynamometer which is overestimating scores) this would be the result of systematic (non-random) error because there is a consistent pattern to the inaccuracy. If there is no consistent pattern to the dispersal measurements scores, this would be considered random error. Systematic error is responsible for bias; random error tends to be self compensating in that errors tend to cancel each other out and thus do not introduce bias.¹⁶ Random error makes inference difficult by obscuring patterns or relationships, which may then necessitate a larger sample in order to reveal one.¹⁶ Therefore, in order to improve validity, attempts must be made to remove systematic error or bias. Conversely, to improve reliability, attempts must be made to remove random error.¹⁶

In the case of measurement reliability the use a protocol designed to maximise the reliability of the method for assessing muscle strength such as standardised testing procedures can reduce measurement error. When we consider validity, we may consider that a dynamometer is a valid test for measuring muscle strength because the transducer responds with a signal proportional to the exerted force. However, in order to infer that the results are valid for the particular individual, we must also look at what the value or score means to the individual being tested; how the data can be interpreted clinically. We must also consider the responsiveness of the method being used. For example, we know MMT uses ordinal data which may not be responsive to small changes in muscle strength whereas the ratio data from dynamometry may be more sensitive to change.

It is essential when looking at the primary research to establish clearly whether a study will provide evidence of the inferential accuracy (validity) of a measurement instrument or its consistency (reliability).¹⁶ Therefore, studies looking at validity or reliability of a measurement, should be evaluated as to their appropriateness for the clinical situations in which the tested instrument is likely to be used.¹⁶

The American Spinal Injury Association (ASIA) first published a standard system for neurological assessment and classification of spinal cord injury in 1982 which was later adopted by the International Medical Society of Paraplegia (IMSOP).⁶ The ASIA classification scale uses MMT in the measurement of muscle strength. However, is MMT the most appropriate assessment method to use or does research support a need for instrumented muscle strength assessment such as HHD?

By reviewing the primary research into the clinimetrics of MMT, ID and HHD in the assessment of muscle strength, physiotherapists and other clinicians can be guided towards using the most reliable, valid, responsive and interpretable method for adults with SCI. A systematic review can therefore answer whether there is support in the research to recommend HHD as the standard tool for measuring muscle strength for adults with spinal cord injury. In addition, gaps in primary research may also be highlighted. A database search did not identify any systematic reviews focusing on the clinimetrics of MMT, ID and HHD in the assessment of muscle strength in individuals with spinal cord

injury. This provides an opportunity for the systematic search, synthesis and summary of the best available evidence in this area.

Statement of the review question

The primary objective of this systematic review was to examine the best available evidence on the clinimetrics of the three main methods of assessing muscle strength in adults with SCI; MMT, ID and HHD. The clinimetric domains included:

- Reliability
- Validity
- Responsiveness
- Interpretability

A further objective was to determine whether it is appropriate for HHD to become the standard tool for measuring muscle strength for adults with SCI.

Overview of the science of evidence synthesis

The ultimate purpose of a profession is to develop a knowledge base that will maximise the effectiveness of practice.¹⁵ Furthermore, the concept of evidence based practice is representative of the fundamental principle that the provision of quality care will depend on the ability of health professionals to make decisions that have been confirmed by sound scientific data and the best evidence currently available.¹⁵

Research methods generally fall into one of two categories; quantitative research which aims to deduce a cause or infer a relationship between variables whilst qualitative research aims to induce meaning or describe a phenomena and its relationships. A systematic review is a research study that systematically identifies relevant primary research studies (either quantitative or qualitative), appraises their methodological quality and summarises their results using a transparent and explicit scientific methodology.¹⁷ Systematic reviews are considered to be the highest level of evidence. They differ from literature reviews due to their rigorous and systematic method guided by a protocol with a detailed description of the review question, methods and criteria used to select and evaluate articles that are to be included. Like a well-designed research study the detail of a systematic review should be such that another researcher could follow the same process and come to the same results.

Systematic reviews are an important source of information for health care professionals. The process of making clinical decisions are complex. In seeking the evidence on which to base clinical practice, clinicians and researchers are faced with a myriad of published papers and web-based sites often offering unclear or conflicting information regarding the choice of intervention approaches, expectations of outcomes or use of measurement tools.¹⁵ Separate research studies which may appear similar in their designs may also reach very different conclusions. Systematic reviews aim to summarise the findings of all relevant individual studies for a particular question, thereby making the available evidence more accessible to health professionals. In addition, a well researched systematic review can provide clinicians and researchers with confidence that up to the time of publication, the information is current and comprehensive on the chosen topic.¹⁴

A meta-analysis was to be conducted if there was an acceptable degree of homogeneity between the included studies. Meta-analysis refers to the statistical analysis of a collection of individual studies to summarise the results from the primary research in order to answer a specific review question or hypothesis.¹⁷ However, a meta-analysis is not an essential part of a systematic review. Meta-analysis can be a very powerful tool for synthesising information across multiple studies, but it can also be misleading if the studies are not appropriately combined. With reference to this review, a meta-analysis was not conducted as the studies were too heterogeneous. Meaning that, for example, their individual research designs, method of assessing muscle strength, the muscle groups assessed were too dissimilar to allow for the integration of the data. Therefore, the results are presented in tables with a narrative summary of the findings.

Methodological issues, critical appraisal, data extraction and data synthesis

The focus of this review has been on the clinimetric properties of the three methods for assessing muscle strength in adults with SCI; MMT, ID and HHD. There has existed a variety in the types of research conducted. Some studies have looked at inter- or intra-tester reliability of one method, some studies have compared methods whereas others have also reviewed the validity of the measurements and the sensitivity to change in test - re-test scenarios. Therefore, with the clinimetric focus it has meant that the studies are not randomised controlled trials or other clear experimental designs that would have allowed for the use of a JBI tool (such as JBI-MASTARI) in the extraction and synthesis of data.

A data extraction form unique to this review was developed in order to extract the important components of each study (see appendix I).

Finding an appropriate tool for assessing methodological quality of included studies proved an interesting process. A systematic review on dynamometry by Stark et al,¹¹ found that the PEDro data collection system provided the best fit for their purposes; however, because the PEDro scale is a tool used to appraise therapy studies, the authors did modify it to suit their review.¹¹ Therefore, their grading form was an amalgam of the modified PEDro tool and the Cochrane Diagnostic Test Accuracy Description.¹¹ The concern of using a similar appraisal tool for this review was regarding the validity of the tool itself as any modification can undermine its own validity and applicability. Therefore, it was decided to use the COSMIN tool to evaluate the methodological quality of the included studies. COSMIN provides a checklist for assessing methodological quality accompanied by a 56 page user manual.¹⁸

The COSMIN checklist was developed in a 2005 International Delphi study of a multi-disciplinary, international collaboration.¹⁹ The COSMIN steering committee consists of 8 members whose backgrounds include clinimetrics, statistics, public health research, health sciences, rehabilitation science and epidemiology, with a number of the committee members being Professors in their field.¹⁸ Although the focus of COSMIN is on health-related patient reported outcomes such as questionnaires, the checklist can also be applied to performance based instruments such as dynamometry and MMT. Over 150 systematic reviews have used the COSMIN checklist. One such systematic review was on the 'clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments', which included muscle strength.²⁰ The authors were able to apply COSMIN to evaluate the methodological quality of their included studies.²⁰ One of the reasons that COSMIN is easy to apply to a wide range of measurement instruments is due to the versatility of the checklist. The checklist is a modular tool containing 10 separate sections including reliability, measurement error, responsiveness, interpretability and various forms of validity including construct and content validity. Therefore, it allows reviewers to use only the applicable section or sections (termed 'box') without compromising the validity of the checklist itself. For this review some studies looked at only reliability so only that box was used whereas others looked at reliability, validity, interpretability and responsiveness, which is reflected in the results section of this review. In addition, there is a box in the checklist regarding generalisability which can be applied to all studies.

The specific definitions of the clinimetric domains used for this review have also followed the COSMIN terminology and definitions which were reached via International consensus. This means that no confusion should be apparent when applying the checklist.

The original checklist was developed with a 56 page manual which contains user friendly, detailed instructions for how to complete the checklist. In addition, background information is provided on the development and validation of the checklist and the rationale behind each item. Possible applications of the COSMIN checklist are described.¹⁸

In 2011, this checklist was further developed to include a 4-point rating scale which can be used to calculate total quality scores (see appendix II). The COSMIN steering committee decided to change the dichotomous response options of the original checklist into 4 response items (excellent, good, fair and poor) in order to increase the discriminative ability of the items.¹⁹ Subsequently a methodological quality score per box can be obtained by taking the lowest rating of any item in the box (“worst score counts”). It is this 4-point scoring checklist which was used for this systematic review.

COSMIN recommend that 2 independent reviewers are used to assess the methodological quality of the included studies with a third reviewer utilised if further resolution is required. The systematic review by Dobson et al.²⁰ had a 3 hour training session on the use of the COSMIN checklist to provide standardisation across the reviewers. This review followed these recommendations with 2 reviewers (and a third as required) as well as a detailed training session prior to evaluating the included studies. The third reviewer was not required during this review as all disagreements were discussed and resolved between the primary and secondary reviewers.

Definition of terms

Spinal Cord Injury (SCI):	Damage either through trauma or disease to the spinal cord which results in a loss of function such as motor control and sensation.
Manual Muscle Test (MMT):	Method of assessing method strength using the Medical Research Council Scale grade 0-5. ⁷
Hand Held Dynamometry (HHD): (also known as myometry)	Portable, hand held method for assessing muscle strength using a transducer.
Isokinetic Dynamometry (ID):	Large, computerised method for assessing muscle strength.
Isotonic contraction:	Muscle tone remains the same throughout the range of movement and should include both concentric and eccentric muscle contractions. ³
Concentric contraction:	The origin and insertion of the muscle's attachments move closer together (i.e. shortening of muscle fibres) during the contraction. ³
Eccentric contraction:	The origin and insertion of the muscle's attachments move further apart (i.e. lengthening of the muscle fibres) during the contraction. ³
Isometric contraction:	The joint angle and muscle length does not alter during the muscle contraction. ³
Isokinetic contraction:	This is a dynamic action where equal speed and resistance is provided by a specialist machine throughout the range of movement as the muscle length changes. ³
Make technique (HHD):	This technique requires the patient to exert a maximal contraction against the HHD, which is in turn resisted by the examiner in a fixed position. ¹²
Break technique (HHD):	This technique requires the examiner to overpower a maximal effort by the patient, thereby producing a

measurement of eccentric muscle strength.¹²

COSMIN:

COnsensus-based Standards for the selection of health status Measurement INstruments is the tool used to assess methodological quality in this review.¹⁸

Reliability:

The degree to which the measurement of muscle strength is free from measurement error. This includes repeated measurement under several conditions such as over time (test-retest), by different people (inter-tester) or by the same person (intra-tester).¹⁴

Validity:

The degree to which the muscle strength test measures the construct it purports to measure. This includes face validity (the degree to which the muscle strength test adequately reflects a measurement of muscle strength) and criterion validity (the degree to which the measurements data is an adequate reflection of a 'gold standard').¹⁴

Responsiveness:

The ability of the muscle strength test to detect muscle strength changes over time.¹⁴

Interpretability:

The degree to which one can assign clinical meaning to the quantitative scores or change in scores in muscle strength.¹⁴

Chapter 2: Methods

The aim of this chapter is to outline and explain the methodology used in undertaking of this systematic review thesis. This review was guided by a systematic review protocol which was published in the JBI library of systematic reviews prior to commencement.²¹ No deviations were made from this protocol.

Review Objective

The primary objective of this systematic review was to examine the best available evidence on the clinimetrics of the three main methods of assessing muscle strength in adults with SCI; MMT, ID and HHD. The clinimetric domains included:

- Reliability
- Validity
- Responsiveness
- Interpretability

A further objective was to determine whether it is appropriate for HHD to become the standard tool for measuring muscle strength for adults with SCI.

Criteria for considering studies for this review

Types of Studies

All quantitative clinimetric studies regardless of design were sought. Studies which reported on the clinimetrics of MMT, ID and/or HHD, as well as those using some form of comparison between the methods used were considered.

Types of participants

Adult population (adult being defined as 15 years and older in reference to the Australian Spinal Injury Registry which categorises young adults as 15 years and older) with SCI, regardless of cause, time since injury, gender or ethnicity.

Focus of review

All studies evaluating MMT, ID and HHD in all of their variations administered by any health professional. However, grip strength dynamometers were excluded from this review due to their specific and specialised nature. The tests also needed to assess muscle strength, any studies looking at muscle tone were excluded.

Outcome Measures

The objective outcomes were measured in terms of the clinimetric domains of reliability, validity, responsiveness and interpretability for MMT, ID and HHD, as previously defined.

Review Methods

Search strategy

A three step search strategy was utilised for this systematic review. The initial search included Cochrane database, Pubmed, CINAHL, SPORTSDiscus, OT seeker, PEDro as well as the specific databases for Australian Physiotherapy Association Journal and The Chartered Society of Physiotherapy Journal as well as the journal Spinal Cord.

The initial search was followed by analysis of the text words contained within the title and subsequently the abstract, as well as the index terms used to describe the study.

A second search using all identified keywords and indexed terms was then commenced across all included databases.

The third step looked at the reference list of all for any additional studies. Only studies published in English were included in this review, with all databases searched from inception to current date (June 2013).

The databases searched were Cochrane database, Pubmed, CINAHL, SPORTSDiscus, OT Seeker, PEDro, Australian Physiotherapy Association Journal, The Chartered Society of Physiotherapy Journal and Spinal Cord.

The search for unpublished studies or grey literature included Mednar.

Initial key words used were 'dynamometry' 'dynamometer' 'manual muscle test' 'spinal cord injury' 'muscle strength'

Additional terms included 'reliability' 'validity' 'responsiveness' 'interpretability' 'clinimetrics' 'adult' 'physiotherapy' 'English'

Details of the search strategy are provided in Appendix III.

Assessment of methodological Quality/ Critical appraisal

The quality evaluation tool COSMIN was used to evaluate the methodological quality of the studies to be included in this review. The COSMIN tool has international acceptance and has been validated via research and therefore, provided a robust approach for assessing clinimetric studies.^{14,18,19}

The COSMIN tool has a 4 point scored checklist (excellent, good, fair, poor) which consists of 10 separate sections, each dealing with a different measurement property. This review will use the domains of reliability, validity, responsiveness and interpretability. The COSMIN checklist is a modular tool which means it is not necessary to complete the whole checklist (see Appendix II). Two reviewers were used during this review process (independent appraisal) and both were educated and familiarised with the COSMIN tool and checklist (education via the 56 page COSMIN handbook).¹⁴ The third reviewer was not required during this review as all disagreements were discussed and resolved between the primary and secondary reviewers.

Data extraction

Quantitative data were extracted from studies using the attached data extraction form which has been developed specifically for this review (appendix I). The data extraction form was developed using inspiration from COSMIN generalisability box^{14,18,19} and the reviews by Stark et al.¹¹ and Dobson et al.²⁰ The purpose of the data extraction form was to highlight all the important data in the included studies from participant characteristics, study design, results and author conclusions in order to compare and contrast the studies during the systematic review process.

Data Synthesis

It was not appropriate to conduct a meta-analysis due to the heterogeneity of the included studies. Therefore, the results are presented in narrative text including raw data presented in the included studies as well as the contextual data. Tables and figures are also used to present the results.

Chapter 3: Results

Chapter 3 aims to present the complete results of this systematic review. The search retrieval process as well as the description of the studies is presented. Appendices IV and V contain the complete list of included and excluded studies. The included studies have formed the basis of the result findings which are detailed in this chapter. Presentation of the results from the included studies, the author's conclusions as well as the assessment of methodological quality and data extraction are also included. The results are presented in tables, figures and narrative descriptive summaries.

Description of studies

A thorough literature search for studies to be included in this review was undertaken between May and July 2013. Education and advice was received from a research librarian leading to the formation of a number of key search terms in order to locate the literature available for this systematic review. The table in Appendix III demonstrates the databases searched including Mednar for grey literature and the results specific to each key term. Limiters were placed on all searches for English and later 'adult', although as this review defines adult as 15 years and older, 'adult' was added later so as not to miss any study which used a younger population. However, all studies located used participants with a much older average age, with a median range of 26.6 to 57 years (see table 4).

The search yielded 2,121 articles; 1,855 after duplicates were removed (refer to Fig 1). The article titles were then screened and those not meeting the inclusion criteria were excluded. Abstracts of the remaining 122 were retrieved and reviewed against the inclusion criteria. This lead to 12 full text articles which were read in detail against the inclusion criteria. Two authors of these articles were contacted via email to ascertain whether they had any research in press or unpublished as their names had appeared on a number of papers. One author did not respond and the other provided the details of an article published in a journal whose electronic database did not date back as far as that particular article. This article was then obtained and included.¹² Two articles were excluded for reasons such as the study examined total motor score rather than individual muscles in the ASIA examination.^{6, 22} The final 11 articles were then reviewed for methodological quality using the COSMIN tool by the primary and secondary reviewers. No disagreements were encountered; therefore the third reviewer was not consulted.²⁴ All 11 were included in the systematic review. Appendix III demonstrates the results of the search as well as the search retrieval flow diagram in figure 1.

The following results are presented in narrative and diagrammatic form as it was not appropriate to conduct a meta-analysis due to the heterogeneity in the results of the included studies. The heterogeneity of the results was due to the differences in design of the included studies rather than an issue with the inclusion criteria of this review. The focus of this systematic review was to look at the clinimetrics for each of the methods for assessing muscle strength. However, only 1 study compared all 3 methods⁴ whereas the other 10 included studies researched only 1 or 2 of the methods.^{12,24,25,26,27,28,29,30,31,32}

The results of the literature search yielded 11 studies to be included for this systematic review. All of the studies used a method for assessing muscle strength in adults with SCI:

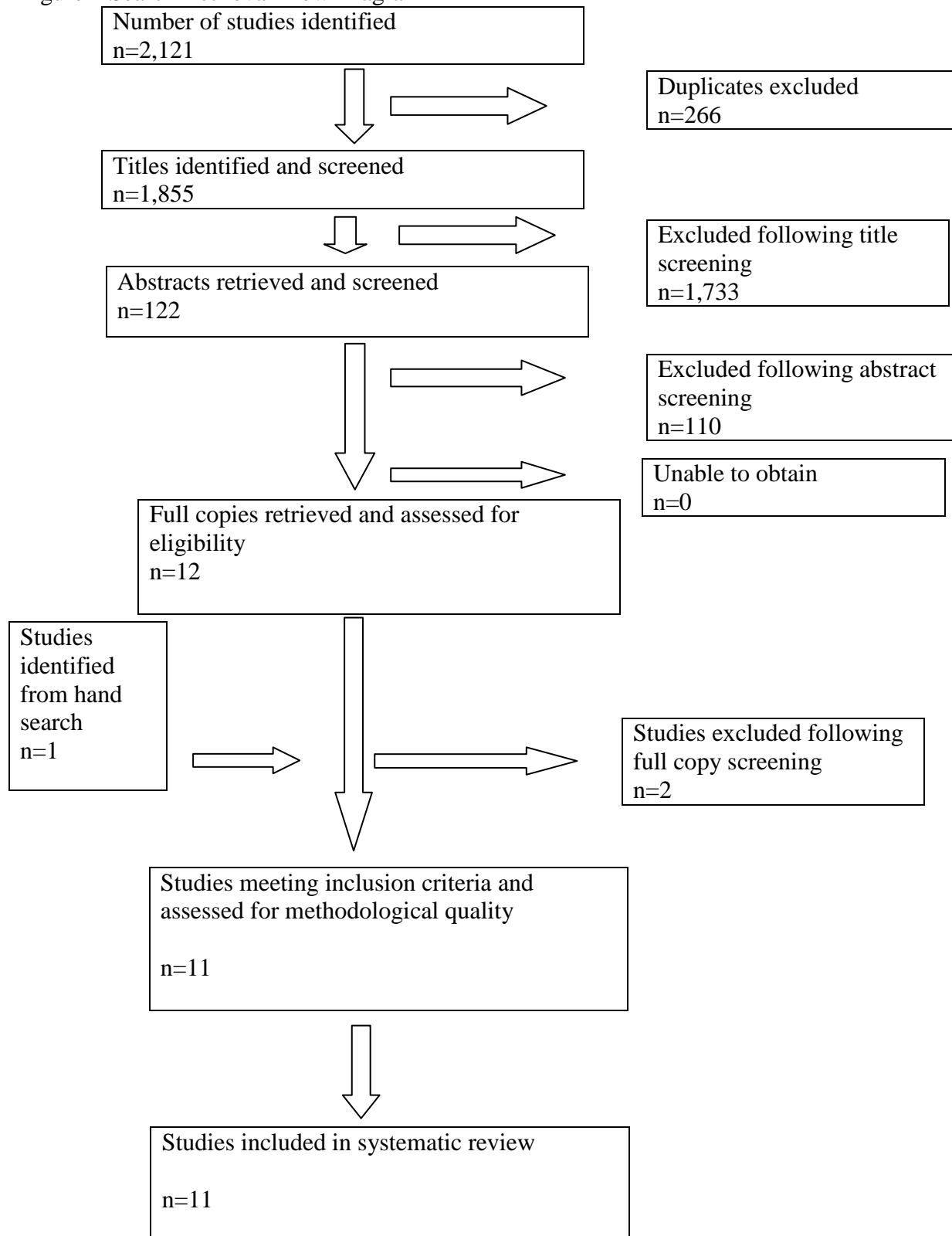
- Two studies used MMT only^{24, 25}
- One study used ID only²⁶
- Three studies looked at HHD only^{12, 27, 29}
- One study compared ID and HHD²⁸
- One study compared MMT, ID and HHD⁴
- Three studies compared MMT with HHD^{30, 31, 32}

Figure 1 shows the search retrieval flow diagram.

Details of the included studies are provided in Appendix IV.

The list of excluded studies (with the reason for exclusion) can be located in Appendix V.

Figure 1 Search Retrieval Flow Diagram



Review findings/results

COSMIN was used to evaluate the methodological quality of the studies to be included in this review as it provided a robust approach for assessing clinimetric studies. It is important to select the most appropriate tool for assessing methodological quality of the studies to be included in a systematic review because this provides the foundation to any results and conclusions that can be drawn from the papers. When studies on measurement properties have good methodological quality, their conclusions are more trustworthy.¹⁶

The COSMIN tool has a 4 point scored checklist (excellent, good, fair, poor) which consists of 10 separate sections, each dealing with a different measurement property. This review used the domains of reliability, validity, responsiveness and interpretability. The COSMIN checklist is a modular tool which means it is not necessary to complete the whole checklist (see Appendix II). By using the 4 point scored checklist rather than the original COSMIN checklist which only had 'yes, no,?, N/A' options has meant that a methodological quality score per box can be obtained by taking the lowest rating of any item in the box ("worst score counts"). A number of papers were scored as 'poor' by the COSMIN tool but the discussion was made between reviewers to include those papers if the 'poor' score related to sample size only. This is consistent with the approach by Dobson et al.²⁰ in their systematic review on clinimetrics. The authors applied the COSMIN checklist to measurement instruments used for the hip, including muscle strength and range of movement instruments, they used this process of next best score on the issue of sample size following discussion with the COSMIN authors.²⁰ For this review, provided the next best score was 'fair' or above on sample size the study was included. The reason this review has decided to also follow this approach is because 7 of the 11 included studies rated poor on sample size which would have masked the other scores given to the study.^{12,24,25,26,27,29,30} In other words, a study would be rated as 'poor' due to a small sample size even when the rest of the study was of a high quality. In addition, for this review it was also decided to include studies which were rated 'poor' for validity if the score related to whether the criterion could be considered a 'gold standard'. ID has been labelled as a 'gold standard'¹¹ but where HHD was compared with MMT, neither are considered 'gold standard' although MMT is usual practice. It was deemed that as long as the next best score was 'fair' or above the study was to be included. This was particularly important as 2 studies were rated 'poor' for validity due to the criterion for comparison but rated much higher for responsiveness and interpretability.^{31,32} However, where the worst score is for sample size it was decided to also list the next best score for that box.

A COSMIN score was given for each of the 11 included studies per domain and then an overall score was given in accordance with the COSMIN user manual. The studies were scored either 'poor', 'fair', 'good' or 'excellent' per domain and the lowest domain score determined the overall study score. The two studies on MMT scored overall 'good',^{24,25} the single study on ID scored 'good',²⁶ the four studies on HHD scored 3 'good',^{12,27,29} and 1 'fair',³⁰ and the four studies that compared methods scored 2 'poor',^{31,32} 1 'fair',⁴ and 1 'good'.²⁸ The 2 studies which scored 'poor' overall was due to a poor score for validity and whether the criterion could be considered a gold standard, MMT is considered usual practice but there is no evidence to suggest that it is gold standard. Otherwise the studies scored higher in the other domains. Others scored 'fair' due to important flaws in study design which included such variables as blinding, standardisation of procedure and

general consistency which will be discussed in chapter 4. Seven of the included studies achieved an overall ‘good’ score^{12,24,25,26,27,28,29} with only 4 scoring below this.^{4,30,31,32}

Table 3 outlines the studies included with particular reference to the clinimetric domain, the score for each domain used and the overall COSMIN score of the study. The full checklist results can be referred to in appendix II.

Table 3: Summary of assessment of methodological quality using COSMIN checklist

Study author/s	Clinimetric Domain evaluated	Study COSMIN score per domain	Overall study COSMIN score
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	Reliability	Poor (sample size) Good	Good
Larson, CA, Tezak, WD, Malley, MS.	Reliability	Poor (sample size) Good	Good
May, LA, Burnham, RS, Steadward, RD.	Reliability, Validity	Reliability Good Validity Good (sample size) Excellent	Good
Schwartz, S, Cohen, ME, Herbison GJ, Shah, A.	Reliability, Validity, Responsiveness, Interpretability	Reliability Good Validity Poor Responsiveness Good Interpretability fair	Poor
Aufsesser, P, Horvat, M, Austin, R.	Reliability	Poor (sample size) fair	Fair
Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF	Validity, Responsiveness	Validity Poor Responsiveness good	Poor
Jonsson, M, Tollback, A, Gonzales, H, Borg, J.	Reliability	Poor (sample size) Good	Good
Kakebeeke, TH, Lechner, HE, Handschin, C	Reliability	Poor (sample size) Good	Good
Noreau, L, Vachon, J.	Validity, Responsiveness, Interpretability	Validity Good Responsiveness Fair Interpretability Good	Fair
Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW.	Reliability	Poor (sample size) Good	Good
Burns, SP, Spanier, DE.	Reliability	Poor (sample size) Good	Good

The completed COSMIN checklists and data extraction table can be located in appendices I and II. Aspects of the data from the data extraction tool are reproduced in the following sections. The results have been divided into the categories of reliability, validity,

responsiveness and interpretability as these domains were the clinimetric focus of the review. However, COSMIN also provide a box called ‘generalisability’ which is in effect a data extraction tool in itself. Although, this box did not have sufficient detail to be used exclusively in this review, it was no less a useful adjunct and therefore, the details have been summarised in table 4.

When comparing the generalisability of the participant specifics, countries and settings above the age of 15 the participant sample for this review represent a broad spectrum with the youngest participant being 15years of age (yrs) and the oldest at 72yrs . The median age ranged from 26.6yrs to 57yrs. All the studies where the distribution of sex is described have males in the majority with 2 studies having male only participants.^{12,32} The described disease characteristics were quite heterogeneous between studies and the reporting varied across papers. Although all studies included SCI, the reporting differed whether the participants were described as traumatic/non-traumatic, classified via ASIA score, paraplegic/tetraplegic or spinal level of injury. The settings vary from in-patient to out-patient settings. The countries included USA, UK, Canada, Sweden and Switzerland. The latter 2 countries do not speak English as a first language; although their studies were printed in English, the language their research was conducted in was not stated.^{24,26} All 11 studies used convenience sampling. The COSMIN tool criteria: ‘percentage of missing items’ was deemed not applicable for this systematic review. This decision was made following detailed review of the COSMIN checklist, manual and a correspondence paper from the authors.²³ The missing items criteria relates to health related patient reported outcomes where questions may be left unanswered or there are missing observations of an item providing an indication that the item was ambiguously formulated, difficult to understand or not relevant to the population. However, as this systematic review looked at quantitative data from a physiological measurement instrument applied by a clinician, it was deemed that this question was not applicable.

Table 4: Summary of Box generalisability from the COSMIN checklist

Authors	Median /mean age (SD)	Distribution Of sex	Disease characteristic	setting	Country And language	Method Used to Select patients	Missing item
Jonsson, M, Tollback, A, Gonzales, H, Borg, J.	Not stated	15M 8F	3 complete 20 incomplete	In-pts hospital	Sweden, Not stated	convenience	N/A
Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW.	40.3 yrs 18-72 yrs	38M 7F	ASIA A-D	In-pt SCI centre	UK, English	convenience	N/A
Kakebeeke,	28-59	Not	T3-12	Out-pt	Switzer-	convenience	N/A

TH, Lechner, HE, Handschin, C	yrs	stated	ASIA A	rehab	land, Not stated		
Burns, SP, Spanier, DE.	57 yrs +/- 15 yrs	19M 1F	ASIA A6, B4, C4, D6	In-pts hospital	USA, English	convenience	N/A
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	53.4 yrs +/- 11.7 yrs	19M	Trauma and non-trauma SCI C4-6 ASIA A,B,D	In-pt SCI ward	USA, English	convenience	N/A
Larson, CA, Tezak, WD, Malley, MS.	32.4 yrs +/- 11 yrs	24M 7F	Para 19 Tetra 19	Out-pt rehab	USA, English	convenience	N/A
Aufmesser, P, Horvat, M, Austin, R.	52 yrs +/- 16 yrs	Not stated but all military veteran	C2-L4 Para 11 Tetra 14	Out-pts, VA Medical centre	USA, English	convenience	N/A
May, LA, Burnham, RS, Steadward, RD.	18-42 yrs 26.6 yrs +/- 6.5	21M 4F	Para 13 Tetra 12	Out-pt clinic	Canada, English	convenience	N/A
Schwartz, S, Cohen, ME, Herbison GJ, Shah, A.	15-70 yrs	122M	Tetraplegic levels C4-6 ASIA A-D	In-pt hospital	USA, English	convenience	N/A
Noreau, L, Vachon, J.	Para 28.2 yrs +/- 13.9 tetra 30.1 yrs +/- 13.4	31M 7F	Para 23 Tetra 15 ASIA A-D	In-pt hospital	Canada, English	convenience	N/A
Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF	15- 68 yrs mean 34 yrs	78M 10F	ASIA A-D	Regiona l SCI centre	USA, English	convenience	N/A

Results

The following results of this systematic review are presented under the sub headings for the clinimetric domains of reliability, validity, responsiveness and interpretability in order to align with the review objective, which was to examine the best available evidence on the clinimetrics of the three main methods of assessing muscle strength; MMT, ID and HHD in adults with SCI. This is preceded by the specific results pertaining to HHD in particular to further meet the review question; to determine whether it is appropriate for HHD to become the standard tool for measuring muscle strength for adults with SCI.

This results section is accompanied by a number of Appendices:

Appendix I data extraction tool

Appendix II COSMIN tool for assessment of methodological quality

Appendix III search results

Appendix IV Included studies

Appendix V Excluded studies

Reliability

In musculoskeletal practice and research, there is often a need to determine the reliability of measurements made by clinicians. Reliability for the purpose of this review has been defined as the degree to which the measurement of muscle strength is free from measurement error. This includes repeated measurement under several conditions such as over time (test-retest), by different people (inter-tester) or by the same person (intra-tester).¹⁴

Eight of the 11 included studies looked at some form of reliability of the method they used to assess muscle strength.^{12,23,24,25,26,27,28,29,30} However, the results varied according to whether they looked at inter-tester or intra-tester or test - re-test data. A number of studies have also gone onto explore the relationships between the methods used which will be discussed under validity. The 2 studies which researched MMT used 10 different muscle groups from the upper and lower limbs determined by spinal level, consistent with ASIA classification.^{24,25} However, the studies focusing on ID and HHD all used upper limb muscle groups^{12,26,27,28,30,31,32} except one HHD study which was the only one to use postural muscle groups.²⁹

The aim of this section is to present in detail the results of the included studies for reliability. However, in order to do this further information is provided regarding the statistical tests that underpin the determination of reliability.

Statistical Tests:

Kappa

The reliability of manual muscle testing, which uses ordinal data, is often calculated using Kappa. The Kappa statistic is a correlated measure of agreement. It indicates the proportion of agreement beyond that expected by chance, that is, the number of achieved beyond chance agreement versus the number of possible beyond chance agreement.³³ However, for some applications, kappa is limited in that it does not differentiate among

disagreements, and assumes that all disagreements are of equal seriousness.¹⁵ Therefore, some researchers use a weighted kappa to specify differential weights for disagreement cells in the agreement matrix as an estimate of reliability. In other words to differentiate disagreement from random differences due to chance or systematic differences due to a consistent pattern in the clinician's data. Kappa ranges from -1 to +1, with zero representing agreement that is no better than expected by chance. However, the interpretation of kappa is not always so straightforward as the magnitude of the coefficient can be influenced by prevalence, bias and non-independence of ratings.³³ If the prevalence of a positive rating is either very high or low, chance agreement is also high and therefore, kappa is reduced.³³ In contrast to prevalence, the effect of bias is greater when kappa is small. The final issue regarding non-independence of ratings is an important consideration for the study designs in this review. An assumption underlying the use of the kappa coefficient is that errors associated with the clinicians' ratings are independent.³³ In studies of intra-tester reliability, knowledge of the first score may influence the second score. Equally when two testers are assessing a patient, the study should be designed in such a way that one tester can not influence the second tester. Therefore, the use of blinding should be administered as far as is practical. Random ordering of patients or using longer time intervals, repeated tests and using a mean instead of single scores and having separate recorders may also influence independence. Where a study appears not to have preserved independence between ratings, kappa should be interpreted with caution.³³

Landis and Koch proposed the following standards for strength of agreement for the kappa coefficient:³³

0.81-1= almost perfect
0.61-0.80= substantial
0.41-0.60= moderate
0.21-0.40 = fair
0.01-0.02 = slight
< 0 = poor

When weighted kappa is used, the magnitude will be affected by the choice of weighting scheme; quadratic or linear weighting.

Intraclass correlation coefficient (ICC)

Both isokinetic and hand held dynamometry use ratio data therefore; most of the studies have used the intraclass correlation coefficient (ICC) as Kappa would be an inappropriate statistical test as discussed earlier. The ICC has several advantages over other correlation coefficients, which is the historical approach to testing reliability. Like other reliability coefficients, the ICC ranges from 0.00-1.00 and it is calculated using variance estimates obtained through an analysis of variance which therefore, means it reflects both degree of correspondence and agreement among ratings.¹⁵ However, the ICC can also be used to assess reliability among two or more ratings, which provides greater clinical applicability as well as not be limited to the same number of testers for each subject, allowing flexibility.¹⁵ In addition, the ICC, although originally designed for use with interval/ ratio data, it can also be applied without distortion of the data on the ordinal scale when intervals between measurements are assumed to be equivalent, therefore, the ICC provides a useful index in a variety of analysis situations.¹⁵

Another major advantage of the ICC is that it supports the generalisability model proposed by Cronbach as a comprehensive estimate of reliability.¹⁵ The concept of generalisability theory is based on the idea that differences between observed scores are due to a variety of factors such as variations in the measurement system including test conditions, alternate forms of a test, characteristic of testers and participants and not just true score variance and random error.¹⁵ Therefore, in the generalisability theory the error variance is multivariant, in that a number of different known and unknown factors contribute to measurement error and overall variance.¹⁵

Portney et al.¹⁵ suggest as a guideline that ICC values above 0.75 are indicative of good reliability, although for many clinical measurements, reliability should exceed 0.90 to ensure reasonable validity. Furthermore, an ICC of 0.95 means that an estimated 95% of the observed score variance is due to true score variance; the balance of the variance (1-ICC=5%) is attributable to error.¹⁷

Confidence Intervals (CI)

Whether using kappa or ICC, the studies should provide a point estimate, which is a single value from their calculations. In addition, given the previously mentioned discussed concern over a single sample value, including a confidence interval (CI) is ideal. A CI provides a range of scores with specific boundaries, or confidence limits that contain the “true” value. The choice of CI depends on the nature of the variables being studies and the researcher’s desired level of accuracy; often the risk of being wrong is reduced by sacrificing precision, therefore, the confidence limits widen as the confidence level increases.¹⁵ In healthcare, the preferred CI are 95% and 99% CI. A 95% CI means that if a study were repeated infinitely then 95% of these intervals would contain the true effect.¹⁷

Manual Muscle Testing:

Two studies researched the inter-tester reliability of MMT, both using the MRC scale grade 0-5.^{24,25} The focus of the 2 studies was on the overall reliability of the International Standards for Neurological Classification of Spinal Cord Injury as developed by ASIA and further endorsed by the International Spinal Cord Society for use worldwide, of which muscle strength testing via MMT is a component. Both studies also reviewed the reliability of sensory testing but this is outside the scope of this review as indeed is the reliability of the ASIA Standards themselves.^{24,25} Both studies used 10 upper and lower limb muscle groups or spinal levels as defined by the ASIA impairment scale (AIS). For ease of interpretation to the other studies, the level and subsequent muscle groups/ actions are listed below:

Spinal level C5 relates to the muscles biceps brachii and brachialis (elbow flexors).
Spinal level C6 relates to the muscle extensor carpi radialis (ECR) (wrist extensor).
Spinal level C7 relates to the muscles triceps brachii (elbow Extensor).
Spinal level C8 relates to the muscle flexor digitorum profundus (finger flexor).
Spinal level T1 relates to the muscle abductor digiti minimi (little finger abductor).
Spinal level L2 relates to the muscle iliopsoas (hip flexor).
Spinal level L3 relates to the muscles quadriceps (knee extensor).
Spinal level L4 relates to the muscles tibialis anterior (dorsi flexor).
Spinal level L5 relates to the muscle extensor hallucis longus (big toe extensor).

Spinal level S1 relates to the muscles gastrocnemius and soleus (plantar flexors).

Table 5 reports on the inter-tester reliability of MMT from the 2 studies.^{24,25}

Jonsson et al.²⁴ describe Kappa as the ratio of the proportion of times that the testers agree (corrected for chance agreement) and the maximum proportion of times that the testers could agree (corrected for chance agreement). They specify the strength of agreement as:

- 0.81-1.00 very good
- 0.61-0.80 good
- 0.41-0.60 moderate
- 0.21-0.40 fair
- < 0.20 poor²⁴

On review of the results from their study 21 out of 40 scores for inter-tester reliability scored greater than 0.61 (good) with 6 of those 21 scores being very good (> 0.81).²⁴ In addition, 13 of the 21 good or very good scores were achieved after the standardisation of assessment versus 8 of the 21 scores before standardisation occurred.²⁴ Which means that 19 of the 40 scores were moderate or below (<0.60).²⁴ The muscle groups which achieved the higher scores and responded well to standardisation included wrist extensors, knee extensors, ankle dorsi flexors and ankle plantar flexors.²⁴ The 4 examiners in this study were considered experienced (2 physicians and 2 physiotherapists).²⁴

Savic et al.²⁵ in their study define the Kappa level of agreement as:

- 0.81-1 almost perfect
- 0.61-0.8 substantial
- 0.41-0.6 moderate
- 0.21-0.4 fair
- 0.01-0.2 slight
- 0 Poor²⁵

Looking at the weighted Kappa results all scores rated almost perfect agreement (> 0.81) except biceps brachii (C5) which was substantial (>0.61).²⁵ However, when reviewing the percentage agreement of the two examiners for the specific MMT MRC scale grades of the left and right sides, grade 0 shows agreement of 97-99%, grade 1 is more variable at 62-80%, grade 2 even less at 50-55%, grade 3 is the lowest score of 22-53%, grade 4 77-83% and grade 5 at 89-93%.²⁵ Therefore, the grades with the highest agreement were grade 0 and 5 and the lowest were grades 2 and 3.²⁵ The examiners in this study were considered experienced (1 clinical scientist and 1 senior physiotherapist).²⁵

Neither study report the kappa CI. Savic et al.²⁵ do state the CI of the ICC when the total motor scores are added together which forms a component of the ASIA Standards. However, these results are not reported here as the focus on this systematic review is on the methods for testing muscle strength not the ASIA Standards and total muscle scores.

In conclusion, in the two studies which looked at MMT, it is apparent that MMT demonstrates some degree of inter-tester reliability.^{24,25} Jonsson et al.²⁴ found reliability increased after standardisation of the assessment procedure and Savic et al.²⁵ found almost perfect inter-tester reliability except for the muscle biceps brachii. However, reliability

was lower for MMT grades 2 and 3.²⁵ Also both studies looked at inter-tester reliability therefore, no comment can be made regarding intra-tester reliability of MMT for adults with SCI.

Table 5: summarises the results of MMT

Author	Specific method used	Muscles assessed	MMT (Kappa and weighted Kappa if stated)					MRC scale grades (% agreement between examiners)
			Inter-tester reliability					
Jonsson, M, Tollback, A, Gonzales, H, Borg, J.	MMT MRC scale	ASIA standards, 10 upper and lower limb muscles (refer to p.36)	Standardisation (kappa)					Not stated
			level	before		after		
				R	L	R	L	
			C5	0.3	high agree	0.57	0.75	
			C6	0.33	0.46	0.67	0.87	
			C7	0.87	0.46	0.47	0.46	
			C8	0.82	0.63	0.57	0.65	
			T1	0.76	0.86	0.62	0.47	
			L2	0.54	0.48	0.56	0.63	
			L3	0.7	0.55	0.7	0.72	
			L4	0.53	0.49	0.89	0.69	
			L5	0.51	0.53	0.66	0.52	
			S1	0.61	0.55	0.64	0.69	
Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW.	MMT MRC scale	ASIA standards, 10 upper and lower limb muscles	level	side	Kappa	weighted Kappa		0: 97-99% 1: 62-80% 2: 50-55% 3: 22-53% 4: 77-83% 5: 89-93%
			C5	R	0.694		0.694	
			C5	L	0.730		0.649	
			C6	R	0.752		0.932	
			C6	L	0.823		0.973	
			C7	R	0.765		0.969	
			C7	L	0.781		0.972	
			C8	R	0.728		0.975	
			C8	L	0.629		0.946	
			T1	R	0.604		0.965	
			T1	L	0.752		0.963	
			L2	R	0.593		0.963	
			L2	L	0.857		0.987	
			L3	R	0.936		0.993	
			L3	L	0.882		0.987	
			L4	R	0.670		0.972	
			L4	L	0.742		0.948	
			L5	R	0.692		0.954	
			L5	L	0.688		0.961	
			S1	R	0.707		0.943	
			S1	L	0.664		0.951	

Abbreviations: L: left and R: right, before and after relates to standardisation of procedure for testing muscle strength

Isokinetic Dynamometry:

Only 1 study researched inter- and intra-tester reliability of ID.²⁶ They used a CYBEX norm II and assessed the muscles related to elbow flexion and extension, shoulder internal and external rotation and shoulder adduction and abduction.²⁶

The researchers performed isokinetic testing of both the concentric and eccentric actions of the muscles at 60 degrees per second as well as isometric contractions.²⁶ The examiners were 2 experienced physiotherapists, although one had more experience with SCI than the other.²⁶ The results for all 3 tests are shown in table 6 below.

Table 6 demonstrates the results for inter- and intra-tester reliability of ID. Kakebeeke et al.²⁶ state in their study that ICC of more than 0.8 can be classified as good. The researchers also calculated a coefficient of variation (CV SD) which is given as a percentage, in order to investigate the agreement between the two examiners. A CVSD of less than 10% was considered reproducible.²⁶

The results show only 4 of the 18 inter-tester reliability ICC scores were >0.8 (good) which were for shoulder internal and external rotation.²⁵ Elbow flexion showed the lowest score at 0.48-0.67. The CV SD% scores for intra-tester reliability show only 2/18 elbow scores to be <10% (reproducible).²⁶ Whereas 26/36 scores for the shoulder were <10% demonstrating better agreement at this joint.²⁶

Therefore, with the results of this study, ID demonstrated varying inter-tester reliability of the shoulder muscles, and poor inter-tester reliability of the elbow muscles.²⁶ Intra-tester reliability of the shoulder was supported but not the elbow.²⁶ The results do not show consistency in whether the ID was more reliable for isometric contractions versus concentric or eccentric movement as the results appear to be more joint than movement dependent. In addition, the results showed inter-tester and intra-tester reliability was more dependent on which joint/ muscle was being tested than the examiner.²⁶

Table 6 summarises the results of ID

Author	Specific method used	Muscles assessed	ID (ICC)				ID (CV SD %)		
			Inter-tester				Intra- tester		
Kakebeeke, TH, Lechner, HE, Handschin, C.	ID CYBEX Norm II	elbow flexors/ extensor, shoulder internal/ external rotators, shoulder abductors/ adductors.	musc	isom	Conc	Ecc	musc	A1/A2	B1/B2
			elb flex	0.56	0.48	0.67	elb flex isom	16.3	12.2
			elb ex	0.62	0.65	0.74	elb flex conc	21.7	15.4
			sh add	0.71	0.73	0.60	elb flex ecc	12.1	12.4
			sh abd	0.74	0.75	0.64	elb ex isom	14.2	8.4
			sh int	0.88	0.88	0.79	elb ex conc	12.2	10.6
			sh ext	0.88	0.74	0.86	elb ex ecc	11.6	15.7
							sh add isom	4.7	6.5
							sh add conc	6.9	4.5
							sh add ecc	7.0	5.9
							sh abd isom	11.1	7.8
							sh add conc	8.6	11.1
							sh add ecc	16.0	8.6
							sh int isom	4.4	6.4
							sh int conc	6.6	5.3
							sh int ecc	7.2	6.2
							sh ext isom	8.7	8.1
							sh ext conc	12.7	13.4
							sh ext ecc	9.4	8.6

Abbreviations: musc, muscle contraction; isom, isometric; conc, concentric; ecc, eccentric; elb, elbow; sh, shoulder; flex, flexors; ex, extensors; add, adductors, abd, abductors; int, internal rotators, ext, external rotators; A1, examiner A 1st measurement; A2 examiner A 2nd measurement; B1, examiner B 1st measurement; B2 examiner B 2nd measurement.

Hand Held Dynamometry:

Five studies researched the reliability of HHD.^{12,27,28,29,30} Two studies used a Chatillon HHD^{12,27}, 2 used a MicroFET HHD^{29,30} and one used a Penny & Giles.²⁸ In addition the muscle actions assessed ranged from 3 studies which used the elbow flexors and extensors,^{12,27,30} with 1 study included wrist extensors,³⁰ and 1 study used shoulder internal and external rotators.²⁸ The final study assessed the postural muscles in a supported and unsupported position.²⁹ Of the 5 studies, 1 study used inexperienced examiners,¹² 2 studies included experienced examiners^{27,28} and 2 studies used a mixture of experienced and inexperienced examiners.^{29,30}

The following table has grouped the results for the inter- and intra-tester reliability for HHD and the muscle assessed in the included studies.

Burns et al.¹² report CI of between 0.86-0.98 for intra-tester reliability using both the 'make' or 'break' technique (with little difference between the two techniques). Inter-tester reliability CI varied from 0.79-0.98 again with little differences between the 2 techniques of 'make' or 'break'.

Burns and Spanier²⁷ report CI of between 0.74-0.99 for the inter-tester reliability of session 1 versus session 2 with the highest value using the break technique at medium and low speeds.

May et al.²⁸ showed correlations for intra-tester reliability of the HHD tests ranging from 0.89-0.96 for shoulder internal rotation and 0.89-0.94 for shoulder external rotation which they attribute to very good reliability. The CI for the paraplegic participants ranged from 0.80-0.96 and for tetraplegic participants 0.81-0.96 for both shoulder internal and external rotation which is very similar in terms of reliability of HHD.²⁸

Larson et al.²⁹ were the only study to look at postural muscles. They compared results when the participants required upper extremity support to maintain an erect posture in sitting for > 5 seconds (supported position) versus those who did not require upper extremity support (unsupported position).²⁹ They also looked at intra- and inter-tester reliability. The intra-tester reliability using ICC ranged from 0.80-0.98 (unsupported) to 0.79-0.99 (supported) for all testers in the 4 directions of force application (anterior, posterior, right lateral and left lateral).²⁸ Inter-tester reliability was even better with ICCs from 0.97-0.99 (unsupported) to 0.96-0.98 (supported) for all directions.²⁹

The final study in the table, Aufsesser et al.³⁰ demonstrated differing results to the earlier studies as their average ICC for inter-tester reliability were much lower and ranged from 0.21-0.84. The muscles they used were similar to Burns et al.¹² study but achieved much lower scores particularly for elbow flexion.³⁰ However, in contrast intra-tester average ICC ranged from 0.93-0.99 which reflected similar findings to the Burns study using the make technique.^{12,30}

Table 7 summarises the results for inter- and intra-tester reliability of HHD

Author	Specific method used	Muscles assessed	HHD, ICC and (95% CI if stated) make or break technique as stated			
			Inter-tester		Intra-tester	
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	HHD Chatillon	Elbow Flexors/ extensors	A make 0.91 (0.79-0.97)	B make 0.94 (0.86-0.98)	A make (1) 0.94 (0.86-0.98)	B make (2) 0.97 (0.93-0.98)
			A break 0.93 (0.82-0.97)	B break 0.94 (0.86-0.98)	A break (1) 0.95 (0.87-0.98)	B break (2) 0.94 (0.86-0.98)
Burns, SP, Spanier, DE	HHD Chatillon	elbow flexors and extensors			Make 0.89 (0.74-0.96)	Low break 0.93 (0.83-0.97)
					Med break 0.97 (0.93-0.99)	High break 0.97 (0.92-0.99)
May, LA, Burnham, RS, Steadward, RD.	HHD Penny & Gilles	shoulder internal / external rotators			musc ext all 0.94 (0.91-0.96)	int all 0.96 (0.94-0.98)
					ext para 0.89 (0.80-0.94)	ext tetra 0.93 (0.86-0.96)
					int para 0.92 (0.86-0.96)	int tetra 0.89 (0.81-0.94)
Larson, CA, Tezak, WD, Malley, MS.	HHD MicroFET	Postural Muscles: Anterior, Posterior, Right lateral, Left lateral	musc anterior (u) 0.98	anterior (s) 0.98	musc anterior (u) 0.87-0.98	anterior (s) 0.87-0.99
			posterior (u) 0.99	posterior (s) 0.98	posterior (u) 0.92-0.97	Posterior (s) 0.83-0.95
			right lateral (u) 0.97	right lateral (s) 0.96	right lateral (u) 0.86-0.97	right lateral (s) 0.85-0.95
			left lateral (u) 0.97	left lateral (s) 0.97	left lateral (u) 0.80-0.93	left lateral (s) 0.84-0.92
Aufsesser, P, Horvat, M, Austin, R.	HHD microFET	elbow flexors, extensors wrist extensors	musc Av sing	elb flex L 0.36 0.22	elb flex R 0.21 0.11	elb ex L 0.89 0.80
			elb ex R 0.74 0.59	wrist ext L 0.84 0.73	wrist ext R 0.84 0.72	Av A sing A Av B
						sing B
						elb flex L 0.93 0.82 0.98
						0.95
						elb flex R 0.99 0.96 0.96
						0.90
						elb ex L 0.98 0.94 0.99
						0.97
						elb ex R 0.96 0.88 0.99
						0.97
						wrist ext L 0.98 0.96 0.99
						0.97
						wrist ext R 0.98 0.94 0.99
						0.97

Abbreviations: A, examiner A; B, examiner B; low, low, low speed of movement; medium, medium speed of movement; high, high speed of movement; musc, muscle; ext, shoulder external rotation; int, shoulder internal rotation; all, results for paraplegia and tetraplegia combined; para, paraplegia; tetra, tetraplegia; u, unsupported position, s:

supported position; Av, average score; sing, single trial; elb, elbow; flex, flexors; ex, extensors.

Two studies were the only ones to compare break/make technique ratio.^{12,27} Burns and Spanier²⁷ also looked at test-re-test data giving a p value. Both studies support that the 'break' technique results in higher strength values than the 'make' technique.^{12,27}

Burns and Spanier²⁷ suggest a p value of less than 0.05 as significant. The 'make' technique and the 'break' technique used at a low speed (angular velocity of 15 degrees/second) resulted in a higher p value, whereas the 'break' technique at medium speed (33 degrees/second) and high speed (55 degrees/second) were considered significant.²⁷

Table 8 summarises the results of HHD for test-retest scenario and break/ make technique ratio

Author	Specific method used	Muscles assessed	HHD test re-test (p value)		Break/Make ratio (B/M) (mean)	
Burns, SP, Spanier, DE.	HHD Chatillon	elbow flexors and extensors	make	0.09	Low B/M	1.16
			low break	0.51	med B/M	1.3
			med break	0.04	high B/M	1.51
			high break	0.007		
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	HHD Chatillon	Elbow Flexors/ extensors			B/M	1.38-1.49

Abbreviations: low, low speed of movement; med, medium speed of movement; high, high speed of movement.

To conclude the results from these studies, it appears that HHD demonstrates reliability for adults with SCI. The 2 studies which used a Chatillon HHD for testing elbow flexors and extensors demonstrated high levels of intra-tester reliability using the 'make' or 'break' technique (ICC > 0.89).^{12,27} Inter-tester reliability was also very good (ICC > 0.91) with the 'break' technique being slightly more reliable than the 'make' technique (ICC > 0.91 versus ICC > 0.93).^{12,27} The study by Aufseßer et al³⁰ used a MicroFET HHD to test the elbow and wrist extensors and demonstrated similar results for intra-test reliability to these studies. However, in contrast they found variable inter-tester reliability at these 2 joints with elbow flexion demonstrated poor reliability (ICC 0.36).³⁰

One study looked at shoulder internal and external rotation using a Penny and Giles HHD. Intra-tester reliability was demonstrated (ICC > 0.89) in this study.²⁸

Only one study looked at postural muscles which demonstrated very good reliability for intra-tester (ICC > 0.96) and inter-tester (ICC > 0.80).²⁹

The two studies that looked at the 'break/ make' ratio used the elbow flexors and extensors and used a Chatillon HHD.^{12,27} They found a 'break/make' ratio of between 1.16 – 1.51 which was dependent on the speed of movement^{12,27} or 1.38-1.49 using isometric contractions.¹² These studies both demonstrate that greater muscle strength is recorded using the 'break' technique versus the make technique for elbow flexors and extensors.^{12,27}

Overall, when comparing the results for reliability for 3 methods of assessing muscle strength; MMT, ID and HHD, there certainly appears to be more evidence demonstrating inter- and intra-tester reliability using HHD compared with MMT and ID.

Validity

The definition of the validity of muscle strength measures used in this review is the degree to which the muscle strength test measures what it purports to measure.¹⁴ This includes face validity (the degree to which the muscle strength test adequately reflects a measurement of muscle strength) and criterion validity (the degree to which the measurements data is an adequate reflection of a 'gold standard').

This section aims to present the results for validity for the 4 studies in this review. The statistical tests used by the different authors are varied. The study by May et al.²⁸ used the ICC to analyse results. However, a number of the other studies used different statistical tests which will be discussed.

Correlation coefficients are used to quantitatively describe the strength and direction of a relationship between 2 variables. Correlation coefficients can take values ranging from -1.00 for a perfect negative relationship, to 0.00 for no correlation, to +1.00 for a perfect positive relationship and the magnitude of the correlation coefficient indicating the strength of association between the 2 variables.¹⁵

Pearson product moment correlation coefficient

This is the most commonly reported measure of correlation, the statistic is given the symbol r for sample data and ρ (rho) for a population parameter. This statistic is appropriate for use when X and Y are continuous variables with underlying normal distributions on the interval or ratio scales.¹⁵

Spearman rank correlation coefficient

Spearman rank correlation coefficient is a non-parametric analog of the Pearson, to be used with ordinal data.

Critical values for the Pearson product moment correlation coefficient and Spearman rank correlation coefficient can be found in a number of resources depending on the studies involved.¹⁴ The observed value for both of these correlation coefficients should be greater than or equal to the tabled value to be considered significant.¹⁴ The stated level of significance will be described with the corresponding studies.

t-test

The study by Herbison et al.³¹ used a t-test to compare the results from MMT and HHD. The t-test allows researchers to evaluate the comparison between two means. Herbison et al.³¹ used the students t test to compare the second with the first HHD measured value and reported this value against the corresponding MMT score for that participant (this was reported as Mean Geometric percent of HHD score 2 as a percent of HHD score 2). P-values were then given.

p-value

The observed p-value is the probability of getting a test statistic as extreme as, or more extreme than, the observed one with the condition that the null hypothesis is true.¹⁷ The

convention is to consider a small p-value as indicating statistical significance of the results; researchers usually check the value of the p-value and if the p-value is less than 0.05 then the results are considered statistically significant.¹⁷

However, Portney et al.¹⁵ suggest that the emphasis placed on significance testing in clinical research must be tempered with an understanding that statistical tests are tools for analysing data and should not be used as a substitute for knowledgeable interpretation of outcomes. It is important to consider the difference that exists between statistical significance and clinical significance as statistical tests are not responsive to the practical or clinical implications of the data.¹⁵

Four of the eleven studies included in this review looked at the correlation between the method to assess muscle strength.^{4,28,31,32} Of the 3 studies that used MMT, all used the 10 point modified MRC scale which includes the addition of 0.5 grades (i.e. 0, 0.5, 1, 1.5 and so on).^{4,31,32} However, Herbison et al.³¹ only used participants with a minimum MMT score of 3.5 or above for participant position consistency. This will be discussed further in chapter 4. Two studies used a CYBEX for ID testing (CYBEX II and CYBEX 340)^{4,28} and all 4 studies chose a Penny & Giles for HHD testing.^{4,28,31,32} All 4 studies also used only upper limb muscle groups with 3 using elbow flexors,^{4,31,32} 1 also using wrist extensors,³² 1 also using shoulder flexors/ extensors, shoulder abductors and adductors⁴ and 1 only using shoulder internal and external rotators.²⁸ All 4 studies used experienced examiners.^{4,28,31,32}

Table 9 summarises the results for validity from 4 studies.

The study by May et al.²⁸ compared data from HHD to ID. ID has been described as the 'gold standard' in a systematic review by Stark et al.¹¹ They used the Pearson product moment correlation to analyse the relationship between the 2 devices. They reported that all coefficients (0.52-0.88) were statistically significant ($p < 0.01$).²⁸ However, the data for tetraplegia for both shoulder external (0.56) and internal rotation (0.52) were significantly lower compared with paraplegia (shoulder external 0.83; shoulder internal 0.74).²⁸

Noreau et al.⁴ were the only study to compare all 3 methods. In relation to the correlation between HHD and ID using the Pearson correlation coefficient ($p < 0.05$ considered significant) for paraplegia 0.70-0.90 and tetraplegia 0.57-0.96.⁴ This is similar to May et al.²⁷, although different shoulder movements were used.⁴ The correlation between HHD and ID was reported by the authors as moderate to strong.⁴

Noreau et al.⁴ also studied the relationship between MMT and HHD. Variable correlations using Pearson correlation coefficient ($p < 0.05$ considered significant) were observed between the strength values measured by MMT and HHD with paraplegia (0.26-0.67) and tetraplegia (0.50-0.95).⁴ Comparisons between adjacent MMT grades were also performed for the strength values measured by ID, paraplegia (0.19-0.65) and tetraplegia (0.35-0.91).⁴ In both these comparisons paraplegia demonstrated the lower scores compared with tetraplegia which differs from May et al, although again different muscles were used.^{4,28}

The study by Schwartz et al.³² compared MMT to HHD. This was the biggest study with 122 participants. They used the Spearman Rank correlation between MMT and HHD,

with 22 of the 24 correlations considered significant at p values less than 0.001 (0.42, 0.59 were the 2 low scores with the other 22 ranging from 0.68-0.94).³²

Herbison et al.³¹ also reported on the relationship between MMT and HHD, although they only used participants with minimum MMT scores of grade 3.5 or higher using the modified MRC scale which has the addition of half grades between the normal MRC scale grading system.³¹ They used the Mean Geometric percent (GM%) changes of strength to show the second measurement obtained by HHD as a percent of the first measurement obtained by HHD (with at least one week in between re-test measurement) and compared these with the results of MMT (which was performed on the same occasions).³¹ The student's t-test was then performed with ranges from <0.001->0.05). The data indicates that changes in scores measured HHD were not always reflected by a change in score by MMT.³¹

Table 9 summarises the results for validity

Author	Methods compared and specific method used	Muscles assessed	Results (statistical test as stated)
May, LA, Burnham, RS, Steadward, RD.	ID CYBEX 340; HHD Penny & Giles	shoulder internal/ external rotators	ICC ext rotn all 0.86 int rotn all 0.88 ext para 0.83 ext tetra 0.56 int para 0.74 int tetra 0.52
Schwartz, S, Cohen, ME, Herbison GJ, Shah, A.	MMT 10 point modified MRC scale; HHD Penny & Giles	elbow flexors, wrist extensors	Spearman Rank Correlation between MMT and HHD time since injury muscle 72h 1w 1m 3m 6m 12m elb flex L 0.86 0.84 0.68 0.82 0.59 0.42 elb flex R 0.80 0.83 0.79 0.68 0.59 0.18 wrist ex L 0.92 0.86 0.81 0.84 0.84 0.77 wrist ex R 0.94 0.78 0.93 0.79 0.75 0.71
Noreau, L, Vachon, J.	MMT 10 point modified MRC scale; HHD Penny & Giles; ID CYBEX II	Elbow flexors/ extensors shoulder abductors/ad ductors shoulder flexors/exten sors	Spearman correlation co-efficients between MMT and HHD para tetra muscles adm d/c adm d/c elb flex 0.48 0.26 0.58 0.48 elb ex 0.46 0.55 0.95 0.88 sh flex 0.63 0.60 0.83 0.50 sh ex 0.44 0.49 0.67 0.57 sh abd 0.64 0.57 0.55 0.59 sh add 0.67 0.34 0.84 0.73 Pearson correlation co-efficients between HHD and ID para tetra muscles adm d/c adm d/c

			elb flex	0.76	0.75	0.81	0.75
			elb ex	0.70	0.82	0.92	0.96
			sh flex	0.89	0.89	0.82	0.78
			sh ex	0.85	0.83	0.59	0.87
			sh abd	0.73	0.82	0.57	0.76
			sh add	0.81	0.90	0.91	0.90
			Pearson correlation co-efficient between MMT and ID				
			Paraplegia 0.19-0.65				
			Tetraplegia 0.35-0.91				
Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF	MMT 10 point modified MRC scale (grade 3.5-5 only); HHD Penny & Giles	elbow flexors	MMT	mean geometric %		p value	
			3.5 =	116		>0.05	
			3.5 to 4.0	205		<0.002	
			3.5 to 4.5	232		<0.05	
			4.0 =	140		<0.05	
			4.0 to 4.5	139		<0.02	
			4.0 to 5.0	191		<0.001	
			4.5 =	127		< 0.01	
			4.5 to 5.0	126		<0.02	

Abbreviations: musc, muscle; ext, shoulder external rotation; int, shoulder internal rotation; all, results for paraplegia and tetraplegia combined; para, paraplegia; tetra, tetraplegia; h, hours; w, week; m, months; adm, admission; d/c, discharge; elb, elbow; sh, shoulder; L, left; R, right; ex, extensors ; flex, flexors, add, adductors; abd, abductors.

When looking at the results for validity, it is more challenging to draw conclusions between the studies due to the heterogeneity of the methods used and statistical tests applied. Of the 3 studies which looked at MMT and HHD all used the 10 point modified MRC scale and a Penny and Giles HHD^{4,31,32}. The 3 studies also included elbow flexors (even though 2 of the studies also looked at other muscles in addition).^{4,31,32} However, one study only used MRC grades higher than 3.5 and the studies differed in statistical tests used (Spearman rank correlation versus Mean Geometric percent).³¹ Good correlation was found for elbow flexors in Schwartz et al study up to 3 months post injury (Spearman >0.68), however correlation declined at 6 and 12 months post injury (Spearman > 0.18).³² Noreau et al.⁴ found more varying correlations for elbow flexors with paraplegics demonstrating lower Spearman's Rank Correlation than tetraplegics (> 0.26; >0.48).⁴ More questionable correlation for MMT and HHD was demonstrated in Herbisons et al.³¹ study which demonstrated that where the HHD score for elbow flexors had changed from test-re-test over time this was not reflected by a change in MMT grade.

Two studies compared ID and HHD. Both used Penny and Giles HHD and both used Cybex ID although different models.^{4,28} Both studies also used shoulder muscles as part of their testing but used different actions. Their statistical tests also differed. The ICC for shoulder internal and external rotation showed good agreement (> 0.86) when both paraplegic and tetraplegic scores were included but on separation of the 2 groups data, paraplegics (ICC >0.74) scored higher than tetraplegics (ICC > 0.52).²⁸ The Pearson Correlation Co-efficient for the second study looked at shoulder abductors/ adductors and shoulder flexors/ extensors.⁴ The results also showed higher agreement for paraplegics (0.73-0.90) versus tetraplegics (0.57-0.91).⁴

Only one study compared the Pearson correlation co-efficient for MMT and ID which demonstrated in contrast lower agreement for paraplegics (0.19-0.65) compared with tetraplegics (0.35-0.91).⁴

Therefore, although one study demonstrated good correlation between MMT and HHD, the other studies show more questionable correlation between these 2 methods, with tetraplegics often scoring better than paraplegics.^{4,28,31}

More consistent correlation was demonstrated in 2 studies comparing ID with HHD and both demonstrated greater correlation for paraplegics versus tetraplegics.^{4,28}

Varied correlation was demonstrated in the one study comparing ID to MMT.⁴

Responsiveness

Responsiveness has been defined for this systematic review as the ability of the muscle strength test to detect muscle strength changes over time.¹⁴

The aim of this section is to summarise the results for responsiveness.

Three studies by Noreau et al.³⁴ Schwartz et al.³¹ and Herbison et al.³² addressed the issue of responsiveness in their research. All three studies used experienced examiners.

Noreau et al reported the significance for all statistical analyses was fixed at the 0.05 level.⁴ For responsiveness they presented their results in diagrammatic form and highlighted the MMT grades which showed significant overlaps with dynamometry. It was reported that there were large variations in the strength measurements for HHD and ID within one grade of MMT, which led to significant overlaps between adjacent MMT grades of each muscle group tested.⁴ This overlap was particularly significant for MMT scores for grade 4 and higher. Less important between grades 3.5 and 4.⁴ Between 3 and 3.5, more distinct differences with smaller overlaps were observed except for shoulder abductors which showed less variability.⁴ The researchers suggest that MMT is not sufficiently sensitive to assess muscle strength, at least in grade 4 and higher as well as in the detection of small or moderate increases of strength over the course of rehabilitation.⁴ Table 10 highlights the muscle groups and the MMT grade which showed significant differences when compared with HHD.

Table 10: muscle groups which showed the largest variation between 2 consecutive grades of MMT when compared with HHD.

Muscle group	Consecutive MMT grades with a p<0.05
Elbow flexors	4.0-4.5 4.5-5.0
Elbow extensors	3.5-4.0 4.0-4.5 4.5-5.0
Shoulder flexors	4.0-4.5 4.5-5.0
Shoulder extensors	4.0-4.5 4.5-5.0
Shoulder abductors	4.5-5.0
Shoulder adductors	4.5-5.0

One of the objectives of Schwartz et al.³² study was to investigate which of the two methods for testing muscle strength (MMT or HHD) best reflected improvement in muscle strength over the course of recovery. They performed a regression analysis on left and right biceps and left and right ECR muscles for MMT and HHD scores over the course of a 24 month recovery time period.³² They used scatter plots and regression lines to represent their data. These scatter plots demonstrated that although MMT data had a smaller increase of scores from time of injury to 24 months post-injury, the HHD data reflect a steady increase in strength as measured in kilograms. Therefore, some of the

small changes in strength were detected by HHD whilst not detected by the MMT.³² In addition, there were inconsistencies demonstrated with the comparison of MMT grade and HHD score.³² Table 11 demonstrates the range of HHD scores measured in kilograms compared with the MMT grade assigned to that muscle group. Left biceps was assigned a MMT grade of 4.5 with a HHD score of 18kg whereas right biceps was assigned a MMT grade of 5.0 with a HHD score of 17.5kg!³²

Table 11 : The range of HHD strength scores in kg compared with the MMT grade for each muscle group

Muscle	MMT range (grade)	HHD range (kg)
Left Biceps	3.5-4.5	5.3-18.0
Right Biceps	3.5-5.0	5.6-17.5
Left ECR	1.5-4.0	2.1-8.8
Right ECR	1.5-3.5	1.9-8.6

Abbreviation ECR: extensor carpi radialis/ wrist extensors; biceps, elbow flexors.

In Herbison et al.³¹ study, which only looked at MMT strength scores of 3.5 and over, they showed that in 3 groups of participants the MMT grade did not change whereas the Geometric Mean percent (GM%) for HHD did (116%-140%).³¹ Large variations of GM% for HHD can be seen for the MMT grades which did change on retesting (see table 9 validity).³¹ In other words, in some instances HHD on retesting showed a change in score which was not reflected in a change of MMT grade which remained unchanged. Suggestive that MMT is not sensitive to small changes in scores in test-retest scenarios, or over the course of the rehabilitation of the patient.³¹

In conclusion, the 3 studies which looked at the responsiveness of dynamometry versus MMT, it was demonstrated that changes in dynamometry scores (ID and HHD) were not always reflected by changes in MMT grades. This suggests that dynamometry using ratio data is more responsiveness to change than MMT which uses ordinal data.^{4,31,32}

Interpretability

Interpretability has been defined as the degree to which one can assign clinical meaning to the quantitative scores or change in scores in muscle strength.¹⁴

As previously stated, there is a difference between statistical and clinical significance. Statistics are essential for interpreting clinical research and it is important to develop a working knowledge of statistics to allow for critical thinking in the use of statistical information for clinical decision making.¹⁴ When looking at interpretability of muscle strength scores, a change of score maybe considered statistical significant but clinically it is not significant because the strength change is too low to affect a change in function.

The aim of this section is to summarise the results for interpretability.

Of the two studies which looked at interpretability, Schwartz et al set out to look at HHD values translated to MMT.³² They used linear regression analyses and scatter plots to show comparison between group MMT scores and HHD measurements from injury to two years post injury. Their data showed that in general, there seemed to be a consistently increasing mean HHD value corresponding to increasing MMT scores from grades 2.5-5.³² However, the range of HHD scores for each MMT grade is much greater, particularly for grades >4.³² Analyses revealed significant differences in the HHD measurements when grouped by the 3 MMT ranges (grades 2.5-3.5, grade 4.0, grades 4.5-5.0) for left and right biceps ($p<0.05$), right ECR ($p<0.1$) during the acute and rehabilitation periods.³² Therefore, there are issues when translating HHD values to MMT grades remembering that MMT uses ordinal data and HHD uses ratio data. Schwartz et al found an established trend supporting the relationship between MMT and HHD.³² However, HHD had a larger range of values and more variability for each corresponding MMT grade as the MMT grade increased towards grade 5, normal.³² In other words, the range of HHD values for a particular MMT grade is more narrow for MMT grades less than grade 4.0 as compared to MMT scores greater than grade 4.0³² (see table 11).

The study by Herbison et al.³¹ also demonstrated that the ranges in HHD scores are far greater than the range in MMT grades (see table 9, validity).³¹ They also showed greater variability in HHD scores (in kgs) as the MMT grade increased towards grade 5.0, normal.³¹ The researchers only used muscle strength data for MMT grades 3.5 or greater but the variation of HHD scores at grade 3.5 was 4.0 kg, grade 4.0 was 5.1kg, grade 4.5 was 6.7kg and grade 5.0 was 9.4kg.³¹

To conclude the results for interpretability, both studies indicate that there was a direct relationship between MMT and HHD in that a positive trend in MMT scores was reflected in a positive trend of HHD scores.^{31,32} However, the range of scores using HHD was far greater than the range of MMT grades.^{31,32} Therefore, it would be problematic to assign a range of HHD scores to a particular MMT grade for adults with SCI based on the results of this research.

Key Findings

Table 12 summarises the key findings from all of the 11 included studies.

Two studies looked at MMT in isolation as part of the ASIA neurological classification examination.^{24,25} Jonsson et al.²⁴ reported that there was weak inter-tester reliability for scoring incomplete SCI, although agreement did improve once the assessment was standardised. However, Savic et al.²⁵ reported very good inter-tester reliability, although reliability varied depending on MMT grade.

Takebeke et al.²⁶ were the only ones to look at ID reliability in relation to SCI. They concluded that ID scores for elbow flexion and extension were not reproducible.²⁶ Isometric contractions of the shoulder showed the most intra-tester reliability which they attributed to the exact positioning of the ID.

A number of studies looked at HHD.^{12,27,28,29,30} Burns et al.¹² in their study reviewing 'make' and 'break' techniques reported high inter- and intra-tester reliability using inexperienced examiners. The study by Burns and Spanier²⁷ went on to further review the relationship between angular velocity and the reliability of strength measures. This was the only study to use HHD during movement rather than an isometric contraction. They reported that high angular velocity during the 'break' technique was associated with higher strength measures.²⁷ They also suggest that strength measures via the break technique are generally higher than via the make technique but maybe less reliable, therefore, the authors conclude that the make technique maybe the preferred option for reliable strength measures clinically.²⁷ Larson et al.²⁹ reported that HHD can be used to objectify postural muscle strength in individuals with SCI, although this was the only study to examine this muscle group. Aufessor et al.³⁰ reported good intra-tester reliability but not inter-tester reliability and recommend fixed arm levers instead of the testers hand to increase reliability.

Four studies looked at a comparison between methods for assessing muscle strength.^{4,28,31,32} May et al.²⁸ reported that HHD measurement of shoulder rotation produced reliable results for paraplegics and tetraplegics but were only comparable to ID testing for paraplegics. The authors suggest the variability of ID scores for tetraplegics maybe related to the method of testing (or modification of the standardised testing protocol maybe required for this patient group).²⁸ Schwartz et al.³² conclude that the range of HHD measurements were less for grades <4 and greater for grades >4. Their results indicated that HHD was able to detect increases in strength over time which were not reflected by changes in MMT grades.³² This last point is echoed by Noreau et al.⁴ who also reported that MMT does not appear to be sufficiently sensitive to determine improvements in muscle strength particularly for muscle strength > grade 4. The authors conclude that HHD is recommended to objectively measure muscle strength.⁴ Herbison et al.³¹ also draw on this point of increases in strength identified by HHD measurements were masked by MMT.

The 11 studies highlight a number of points which will be further explored in chapter 4: Discussion, namely, standardisation of testing procedure, make v break techniques, modification of testing protocols specific to SCI and in particular tetraplegia, the use of fixed lever arm rather than the testers arm when using HHD and the lack of sensitivity of

MMT particularly > grade 4.

Table 12: summaries the Authors key findings/ conclusions

Study	Method assessed	Authors conclusions
Jonsson, M, Tollback, A, Gonzales, H, Borg, J.	MMT	weak inter-rater reliability for scoring incomplete SCI. Agreement improved for motor scores once assessment standardised
Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW.	MMT	very good levels of agreement between 2 experienced examiners in all components of ASIA neuro exam
Kakebeeke, TH, Lechner, HE, Handschin, C	ID	In subjects with SCI isokinetic strength scores of elbow flexion and extension are not reproducible. Isometric contractions of the shoulder show most intra-tester reliability as the exact repositioning of the ID was guaranteed
Burns, SP, Spanier, DE.	HHD	Findings confirmed the authors hypothesis that greater angular velocity during break technique HHD testing is associated with higher strength measures. Variability in angular velocity is a potential source of error. To minimise variability in strength measurements maybe better to use make technique
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	HHD	make and break techniques show high inter- and intra-tester reliability when performed by inexperienced examiners
Larson, CA, Tezak, WD, Malley, MS.	HHD	HHD can be used among different examiners to objectively quantify postural muscle strength in sitting for individuals with SCI.
Aufsesser, P, Horvat, M, Austin, R.	HHD	intra-tester showed good reliability whereas inter-tester was not reliable or practical in a clinical setting. Authors recommend fixed lever arm instead of testers hand to increase reliability
May, LA, Burnham, RS, Steadward, RD.	ID and HHD	HHD measurement of shoulder rotation in SCI produced reliable results for paraplegics and tetraplegics and were comparable to ID testing for paraplegics. The variability of the isokinetic scores in tetraplegia indicates that this may be an issue with the method of testing (?modified

		testing protocol)
Schwartz, S, Cohen, ME, Herbison GJ, Shah, A.	MMT and HHD	Range of HHD measurements for a particular MMT grade were better for grades <4 and less specific for grades < 4. Results indicate that HHD detect increases in strength over time which are not reflected by changes in MMT scores
Noreau, L, Vachon, J.	MMT, ID and HHD	MMT does not appear to be sufficiently sensitive to determine improvements in muscle strength especially >grade 4. Recommend HHD to objectively measure muscle strength
Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF	MMT and HHD	results of study suggest that increases in strength identified by myometric measurements were masked by MMT

In conclusion, the results of the 11 studies included in this systematic review examining the clinimetric properties of MMT, ID and HHD have been presented. All studies used participants with SCI but the studies were otherwise heterogeneous in terms of which method used, which muscle groups were assessed and which clinimetric domain was the focus of the study. Therefore, despite being unable to conduct a meta-analysis a number of important issues have been highlighted which bear relevance to the clinical application of the 3 methods for assessing muscle strength which will be further explored in Chapter 4, discussion.

Hand Held Dynamometry

A further objective of this systematic review was to determine whether it was appropriate for HHD to become the standard tool for measuring muscle strength for adults with spinal cord injury.

Further summary of the evidence is required to answer this (see table 13).

Eight of the 11 studies assessed HHD in various forms in their research.^{4,12,28,29,30,31,32} With regard to reliability, six studies reported some degree of reliability for HHD,^{4,12,27,28,29,30} as well as Schwartz et al.³² who established the reliability of HHD pre-study. Three studies reported both intra-tester and inter-tester reliability.^{12,28,29} There were conflicting views whether HHD was inter-tester reliable, with one study reporting intra-tester reliability but not inter-tester.³⁰ However, all the studies which examined intra-tester reliability demonstrated a positive reliability.^{4,12,27,28,29,30} Studies examining HHD produced the most amount of data, therefore, the lack of data for the other methods may confound the findings. However, there appears to be support in the literature that HHD is reliable.^{4,12,27,28,29,30}

Four studies assessed the validity of HHD in their research.^{4,28,31,32} All four concluded that there was positive support for the validity of HHD, in that HHD was able to measure the strength of a muscle and changes in muscle strength over time.^{4,28,31,32} HHD was also comparable to the ID in the two studies which compared the two methods.^{4,28}

Of the 3 studies which looked at responsiveness, these studies supported that HHD is responsive to changes in muscle strength especially when compared with MMT.^{4,31,32} The 3 studies all reported that HHD was able to demonstrate small changes in strength scores which were not always detected by MMT.^{4,31,32}

Two studies discussed interpretability particularly in comparison to MMT.^{31,32} Due to the ability of HHD to detect small changes in muscle strength this meant that the results were able to be interpreted clinically as to whether a participant had improved scores even though MMT grades were unchanged.^{31,32} Translation of HHD scores to MMT grades is highly questionable.^{31,32}

In comparison, the research on MMT in relation to SCI has shown varied inter-tester reliability depending on the grade of MMT, questionable validity, lack of responsiveness and questionable interpretability.^{4,23,24,31,32}

ID was reported with varied reliability depending on joint/ muscle action and correlated well with HHD, although more so with paraplegics compared with tetraplegics.^{4,26,28}

Table 13: summarises the results for each method of assessing muscle strength against the clinimetric domain used in the study,

Authors	method	Reliable?	Valid?	Responsive?	Interpretable?
Jonsson, M, Tollback, A, Gonzales, H, Borg, J.	MMT	Poor for inter-tester but improved once procedure standardised			
Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW.	MMT	Yes grades 0 and 5 varied grades 2,3,4			
Kakebeeke, TH, Lechner, HE, Handschin, C	ID	No with elbow movt. Yes isometric shoulder movt			
Burns, SP, Spanier, DE.	HHD	Yes at specific speed of movement			
Burns, SP, Breuninger, A, Kaplan, C, Marin H.	HHD	Yes inter and intra- tester			
Larson, CA, Tezak, WD, Malley, MS.	HHD	Yes intra- and inter- tester			
Aufsesser, P, Horvat, M, Austin, R.	HHD	Yes intra- tester No inter- tester			
May, LA, Burnham, RS, Steadward, RD.	ID	yes			
	HHD	Yes	yes– good relationship between ID and HHD		
Schwartz, S, Cohen, ME, Herbison GJ, Shah, A.	MMT		questionable	no	Possibly < grade 4, No> grade 4
	HHD	Yes pre- study	yes	yes	yes

Noreau, L, Vachon, J.	MMT		No compared with HHD	No compared with HHD	
	ID		Yes compared with HHD	yes	
	HHD		Yes compared with ID	yes	
Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF	MMT		questionable	no	questionable
	HHD		yes	yes	yes

Abbreviations: movt, movement.

However, a number of studies do raise a number of issues relating to the application of HHD for assessing muscle strength in adults with SCI. These include whether the make or break technique is used, the strength of the examiner/ tester, whether a fixed lever arm is used, isometric contractions versus movement for testing, the experience of the examiner, testing procedures. These issues will be further explored in Chapter 4, discussion.

In conclusion, the results of this systematic review show that there is support in the literature for the use of HHD in the assessment of muscle strength in adults with SCI.^{4,12,27,28,29,30,31,32} The results appear to suggest greater reliability, validity, responsiveness and interpretability for HHD compared with MMT and ID, although this must be viewed in context of the small number of studies used for this systematic review, the muscles used and the heterogeneity of study design and level of credibility of the included studies.

Chapter 4: Discussion

This chapter aims to discuss the results of the included studies as well as further explore the author's conclusions and relate these to other research findings. The discussion will be organised under a number of sub-headings for ease of reference to the results as presented earlier, including a separate subheading for HHD to further discuss the second objective for this review.

Methodological Quality

The first point to discuss is methodological quality of the included studies. The COSMIN tool was used for this systematic review. A score was given for each of the 11 included studies per domain and then an overall score was given in accordance with the COSMIN user manual.¹⁸ The studies were scored either 'poor', 'fair', 'good' or 'excellent' per domain and the lowest domain score determined the overall study score. Seven of the included studies achieved an overall good score^{12,24,25,26,27,28,29} with only 4 scoring below this.^{4,30,31,32} It was deemed appropriate during protocol development to include all studies despite methodological quality due to the small number of studies and their heterogeneity.²¹ A meta-analysis may have been able to give weight to the better quality studies, however, as this was not possible, it is hoped that through the narrative summary that the quality issue is addressed. The subject of scoring could also be indicative of the methodological quality assessment tool used. COSMIN was deemed the best fit for this review but a more specific tool on the clinimetric properties of measurement tools/devices may need to be developed for future systematic reviews.

Clinimetrics

SCI, either traumatic or non-traumatic in origin, and the disabilities that follow are of great concern to the medical world, to the person affected, their family and society as a whole.³⁴ Testing of muscle strength in this patient group is necessary in order to classify the neurological impact of the injury, treatment planning and outcome evaluation regardless of whether MMT, HHD or ID are used. In addition, if the measurement of muscle strength is to be utilised by the clinician and bears relevance to clinical practice and decision making, then it must fulfil at least four criteria: firstly it must be sensitive enough to distinguish sub normal from normal muscle strength; second, it must be precise enough to detect both increases and decreases in strength; third, it must be reliable; and lastly it should be predictive of other variables of potentially greater importance.³⁵

The International College of Applied Kinesiology established an operational definition for the use of MMT although it can be applied to any form of muscle strength assessment: "MMT evaluates the ability of the nervous system to adapt the muscle to meet the changing pressure of the examiner's test. This requires that the examiner be trained in the anatomy, physiology and the neurology of the muscle function. The action of the muscle being tested, as well as the role of synergistic muscles must be understood. MMT is both a science and an art."(p.27)³⁶ MMT has long been the standard method in the measurement of muscle strength. MMT is simple, easy to use, and therefore clinically practical.¹⁰ However, although there is evidence to support that MMT is a reliable

method and has been shown to predict specific or general function in some patient populations, it has also been shown to be neither precise nor sensitive.³⁵ HHD potentially quantifies strength numerically by recording force output and because of the portability of HHDs, they are useful in the clinical/bedside environments for tracking strength changes.¹⁰ The reliability of HHD, it has been suggested is highly dependent on standardized testing procedures, particularly that the examiner has sufficient strength to resist the subject's muscle contraction; the strength of the examiner should be greater than that of the specific muscle group being assessed.¹⁰ A drawback of HHD is that strength is often tested at one point in the ROM,¹⁰ although the study by Burns and Spanier²⁷ overcome this limitation by using movement. However, this then may present its own issues related to the standardisation of speed and indeed its reproducibility on retesting. ID allow for the measurement of force/torque throughout the available ROM, both concentrically and eccentrically. These devices are not portable and, therefore, not suitable for quick or bedside clinical tests.¹⁰ They are also expensive and can take time to set up and calibrate and therefore, not always readily available to the clinician.

Measuring strength changes in subjects with SCI can provide evidence of recovery or deterioration of motor output and thereby provide diagnostic information as well as direct the efficiency of rehabilitation interventions.

In modern medicine the importance of clinimetric issues is increasing for several reasons. First, the strong emphasis on evidence based medicine requires that the value of clinical interventions to be shown empirically, which implies measurement of outcomes.³⁷ Second, because of the current high standards of medicine in the Western world, the added value of diagnostic tests and therapeutic interventions become more scrutinised.³⁷ So, if we still want to show improvements, our measurement instruments must meet the challenge to detect smaller and smaller changes.³⁷

Eleven studies were identified as meeting the inclusion criteria for this systematic review. Of the 11 studies, 8 examined reliability,^{12,24,25,26,27,28,29,30} 4 examined validity,^{4,28,31,32} 3 examined responsiveness^{4,31,32} and only 2 examined interpretability.^{31,32} In addition, 5 studies included MMT,^{4,24,25,31,32} 3 included ID^{4,26,28} and 8 included HHD.^{4,12,27,28,29,30,31,32} Only one study reviewed all 3 methods.⁴

When reviewing the results it is useful to bear in mind, the context of each of the 3 methods for assessing muscle strength. MMT is the most commonly used method and could be considered as the standard practice tool, particularly as this is the recommended tool for assessing SCI using the ASIA standards. However, ID has been reported as the gold standard for assessing muscle strength¹¹ but this is perhaps questionable in SCI.²⁸ HHD is both portable and quantifiable.

Reliability

Two studies looked at the reliability of MMT.^{24,25} However, the focus of their research was on the reliability of the motor and sensory examination performed according to the ASIA standards, of which MMT is an important component. The sixth edition of the International Standards for Neurological Classification of SCI was revised in 2000. These standards were developed by ASIA for assessing and classifying the neurological deficit in patients with SCI and it is known as the ASIA Impairment Scale, AIS (see table

1). These standards have been endorsed by the International Spinal Cord Society and are used clinically and for research purposes worldwide.²⁴ The AIS uses a motor and sensory examination and the results of these are used to give an overall grade (A-E, where E is normal). The motor examination uses the key upper limb and lower limb muscles innervated at 10 spinal levels (C5-T1 and L2-S1, see p.35). The strength of these muscles is assessed using MMT on the MRC scale (grades 0-5).⁷ Therefore, individuals scores are given for each muscle but also a total score out of 50 is given for the left and right sides. Classification is then determined whether there is sacral segment preservation, (motor and/or sensory) and whether motor function is preserved in <50% or >50% of key muscles (grade 3 or higher on the MRC scale) below the neurological level of injury. Therefore, the sensitivity of the AIS could be considered decreased because in order to be classified as a grade D incomplete, >50% of the muscles below the neurological level of injury need to score grade 3 or higher on the MRC scale. In other words, whether the muscle scores a grade 3, 4 or 5 is not particularly relevant for classification using the AIS as they are considered the same, although the total score will show differences when the individual grades are added together for recording purposes. The 2 included studies by Jonsson et al.²⁴ and Savic et al.²⁵ looked at the inter-tester reliability of the individual MMT for each of the 10 muscles groups recorded in the AIS, which is why it was included in this review. There have been numerous other studies which have then looked at the reliability of the total scores and grade classification of the AIS which is quite different and was the reason 2 further papers were excluded.^{6,22} Savic et al.²⁵ did find poor inter-tester reliability of grades 2 and 3 which they state can have implications on the AIS as previously reported.²⁵ However, these differences did not influence the overall AIS grade classification of the patients in their study. As previously mentioned it is not the purpose of this review to examine the reliability of the ASIA Standards or AIS as this is a separate measurement tool. The clinimetric properties of MMT may have little bearing on the reliability and validity of the AIS as a whole because it uses total motor scores.

Jonsson et al.²⁴ reported varied inter-tester reliability of MMT but improvements were seen after standardisation of procedure. However, this point of standardisation is not unique to MMT. When considering the repeatability of a test standardisation and well documented testing procedure or protocol is important. The International College of Applied Kinesiology also make this point and state that the following factors must be carefully considered when testing muscles in clinical and research settings.³⁶

- Proper positioning: test muscle should be the prime mover
- Adequate stabilisation of the regional anatomy (particularly the proximal segments i.e. it maybe important to stabilise the trunk and shoulder in order to test the elbow)
- Observation of how the patient maintains the test position
- Observation of the patient during performance of the test
- Consistent timing, pressure and position
- Avoidance of preconceived impressions regarding the test outcome
- Non-painful contacts – non-painful execution of the test

- Contraindications due to age, debilitating disease, acute pain and local pathology or inflammation.³⁶

In the included studies Kakebeeke et al.²⁶ found issues with the reliability of elbow flexion and extension which they attributed to the difficulty in attaining exact positioning of the ID at this joint. May et al.²⁸ also state that exact positioning may have been a factor in their research with their tetraplegic participants. In the absence of adequate stabilisation, the measurement of strength may be over or under estimated. Stabilisation should be used to prevent substitution or force augmentation by muscle groups other than those that are the focus of measurement.³⁵ It is possible that ID can not be configured properly for clinically relevant muscles which were not the subject of testing in the included studies (such as the more distal muscles or lower limb). Shoulder flexion and abduction strength tests can be problematic because the fixed axis can become misaligned with the changing axis of rotation of the glenohumeral joint.¹⁰ Noreau et al.⁴ in their study tested shoulder abduction/adduction with the trunk semi-reclined at 45 degrees to mitigate this as well as adding trunk stabilisation. In order to provide adequate stabilisation it is also important that the patient can maintain this position for the full duration of the test/s. Consistent timing, pressure and positioning are important whether using MMT or dynamometry. The time given should be sufficient for the muscle to achieve a maximal contraction; not too fast so the contraction is submaximal or too slow that the muscle fatigues. This goes in tandem with pressure; the pressure should not be too great and cause discomfort or fatigue the muscle at a submaximal level but not so light as to lose control of the muscle whereby producing uncontrolled submaximal movement. Positioning of the resistance or device is also important when considering leverage and should be well documented.

Avoidance of preconceived outcomes is applicable to MMT and dynamometry. The examiner needs to be mindful of verbal and non-verbal cues that may adversely affect the test.

The measurement of muscle strength should be non-painful and the examiner needs to understand when to abort a test due to pain or indeed muscle spasm and to document this. It is vital when using dynamometry that the placement of the device does not elicit pain as this will result in muscle inhibition and sub maximal scores.

Understanding contraindications are as important for muscle strength assessment as they are for any other assessment or clinical procedure.

Whichever method of assessing muscle strength is used, a number of factors, both intrinsic and extrinsic, as mentioned above must be considered and always clearly documented in order to increase the repeatability and therefore reliability of the test on subsequent occasions.

Validity and Responsiveness

Validity and responsiveness will be discussed together as it can be argued that the lack of responsiveness of a device is the reason for its questionable validity. Of the 2 studies which examined the reliability of MMT a number of issues were raised, such the varying reliability of the different MRC scale grades and the use of standardised procedures.^{24,25}

As discussed in the introduction an instrument could be regarded reliable but not necessarily valid. Although MMT did demonstrate some degree of inter-tester reliability in the included studies, a number of studies which then compared MMT to ID or HHD found overlapping results and a much greater range of dynamometry scores within one MMT grade which puts in doubt the validity of MMT, generally through the lack of responsiveness of the MRC scale. This issue of overlapping scores between MMT and HHD is supported in other studies using different participant populations.³⁵ In addition it has been reported that differences in strength of 20-25% were not detected by MMT.³⁵ Therefore, when starting to interpret the results of MMT there are issues related to a muscle which may have increased in strength but not enough to move up a whole grade on the MRC scale and therefore, can not be quantified. This is particularly relevant at grades >4 as mentioned by a number of authors.^{4,12,32} Of the two studies which examined the reliability of MMT, Savic et al.²⁵ reported lower reliability scores for grades 2,3,4. They go on to state that this can have implications on the ASIA impairment classification and could in some cases result in classifying the same injury as ASIA grade C by one examiner's results and ASIA grade D by another.²⁵ However, the authors do not go on to discuss the possible explanation for poorer reliability at these grades.²⁵ Grade 4 (movement against gravity and resistance) and grade 5 (normal) were the grades most reported to be unreliable in a number of the included studies.^{4,12,31,32} This maybe due to the subjective nature of the amount of resistance to apply in order to classify these grades which may differ from examiner to examiner.³⁸ However, Savic et al.²⁵ report high reliability for grade 5. This maybe partly due to the purpose of their research which was to look at the reliability of the motor examination as part of the ASIA standards.²⁵ Therefore, the examiners in their research may have been satisfied for the purposes of ASIA classification that the muscle/ neurological level being tested was 'normal'. However, in other studies especially those comparing methods of measuring muscle strength encounter issues with grades 4 and 5 with less of an issue with grades 2 and 3 which uses gravity as a standard.³⁸ A number of the included studies in this review reported greater variations of overlapping of scores for grades 4 and 5.^{4,32} Therefore, a participant may have an elbow flexion score of 18kg via dynamometry testing and be classified as grade 4.5, whereas another participant may have a score of 17.5kg and be given a grade 5.³² The issue particularly with grade 5 is that there is a broad range of scores for 'normal'. In other words normal elbow flexor muscle strength of a 90 year old lady would be vastly different on dynamometry testing compared with a 24 year old rugby player. Therefore, quantifying normal can be problematic using MMT. A participant may be given a grade 4 score because the muscle is able to act against gravity and some resistance but is not considered normal because the resistance was considered low by the examiner. Testing the muscle on the other side may then gain a grade 5 because it was able to withstand greater resistance than the first. Is this scoring correct or is the subject naturally stronger on one side? Herbison et al.³¹ develop this point further and state that in SCI strength tested as grade 5 may mask a relative decrease in strength compared to the pre-injury status. A subject maybe strong enough to achieve a grade 5 when in fact strength was less than normal compared to the pre-SCI status of the muscle. In addition, once someone has achieved a grade 5 then no further improvement can be recorded even if the examiner detects increases in strength.

The point of sensitivity and therefore responsiveness is made by Schwartz et al,³² MMT

grades 4 and 5 which apparently fail to detect improvements in strength that are reflected by increasing dynamometry scores for a given muscle during recovery. This lack of sensitivity of MMT is not isolated to SCI. A study on Parkinsonian patients found 17 complained of muscle weakness, but all were tested as normal by MMT.³⁵

The correlation between methods is another worthy discussion point. Schwartz et al.³² found good correlation between MMT and HHD until 6 months post injury where a decline was seen. The lower correlations for biceps in particular at 12 months could be explained by the truncated range of MMT scores.³² The majority of patients in this study had a MMT grade of 4 or 5 at 12 months which showed a wide range of dynamometry scores.³² It appears that as MMT scores enter the good to normal range, they become less sensitive to detecting subtle increases in strength as measured by HHD.³² This lack of sensitivity of MMT was also the explanation given by Noreau et al for the declining correlation between MMT and HHD from admission to discharge for the tetraplegic participants.⁴ The strength of the participants improved over the course of the rehabilitation which was observed with HHD but not MMT.⁴ The correlation for tetraplegics was generally better than for paraplegics for MMT and HHD which has been explained by the weaker scores for this group of participants except towards discharge and the fact that paraplegics scored more highly in the grade 4 to 5 range as discussed earlier.⁴ Correlation between ID and HHD was also varied in 2 studies.^{4,28} In contrast correlation of ID and HHD was better for paraplegics versus tetraplegics.²⁸ The lowest correlation observed in one study was for shoulder extensors and abductors which the author attributed to execution problems using ID for individuals with tetraplegia which limits their capabilities to produce maximal effort.²⁸ Joint proximal stability may be compromised leading to an inability to produce real maximal muscle contractions due to shoulder girdle instability, lack of trunk stability, joint laxity or muscle imbalances.²⁸ May et al.²⁸ question ID to be labelled as a gold standard in SCI due to these inherent issues with ID positioning which is supported in the included studies.

Therefore, of the studies which looked at validity and responsiveness, the validity of MMT is questioned due to its inability to detect subtle changes in muscle strength particularly at the grades of 4 and 5.^{4,12,31} Correlation between MMT and dynamometry showed significant overlapping of scores particularly in the good to normal range.^{4,31,32} Correlation between ID and HHD was more favourable but varied depending on the muscle action and with positioning issues for ID with tetraplegia.^{4,28}

Interpretability

It has been said that as tools, the therapist's hands are the most sensitive, fine tuned instruments available and that the best instrument to perform MMT is a well trained examiner, using his perception of time and force with knowledge of anatomy and physiology of muscle testing.³⁶ However, the therapist can still review a patient's movement quality using MMT as part of the examination but add dynamometry to objectify it and therefore, make the results more meaningful and easier to interpret. The results of HHD are often recorded in kilograms and ID in Newtons which are measurement scales of equal interval which allow comparison between individuals. This is perhaps easier to understand than a grade of movement using MMT with unequal

intervals between grades and an inability to compare strength between individuals as the 90 year old lady's 'normal' is not the same as the 24 year old rugby male's 'normal'.

Additional variables

An area worthy of discussion is related to type of movement assessed. MMT often looks to achieve concentric movement at the lower grades and isometric movement at grades 4 and 5 once resistance is required. ID has the advantage over MMT and HHD because it is able to assess all types of movement including concentric, eccentric, isometric and isokinetic contractions. Which means that it can assess both static and dynamic movement. Traditionally HHD used only isometric contractions using either a make or break technique. However, as technology improves, HHD are becoming increasingly sophisticated to allow for dynamic movement.²⁷ However, whatever type of contraction used in the assessment procedure regardless of method used, reliability and validity of the test is compromised if used interchangeably therefore, testing procedure should be well documented.

Another testing variable was blinding of examiners and/ or participants. Of the 11 included studies, 5 were not clear on the issue of blinding,^{4,12,30,31,32} 2 stated that no blinding was present,^{25,27} 1 blinded participants only²⁶ and 3 blinded both.^{24,28,29} Some authors stated that it was hard to blind examiners to the measurement scores which could certainly be said of MMT when the examiner is responsible for determining the grade and with some HHD where the maximum score is not held in the memory of the device but needs to be observed by the examiner.^{25,27} However, blinding is considered an important source of bias for both participants and examiners. Knowledge of previous scores can directly influence subsequent scores. Jonsson et al.²⁴ who looked at inter-tester reliability of MMT blinded each examiner to the other examiners results, which showed poor inter-tester reliability. May et al.²⁸ compared ID and HHD with both participants and examiners blinded which showed good reliability and validity for these devices. Larson et al.²⁹ also used blinding for HHD for postural muscles which showed good reliability. Kakebeeke et al.²⁶ blinded participants only and showed mixed reliability results. Whether blinding was used did affect the methodological quality scores as assessed by COSMIN and therefore, reflects on the overall quality of the study and therefore, any results and conclusions drawn. Therefore, blinding is an important factor and should be considered relevant by all researchers.

Verbal encouragement also varied between studies. Providing verbal encouragement toward maximal effort may influence the scores. However, whether verbal encouragement was utilised in each particular study may not be relevant as the scores of each patient are not being directly compared. What is important is that consistency is given in test conditions particularly when 2 or more methods were compared. Of the studies which used verbal encouragement this remained consistent throughout the research.^{26,28} Noreau et al.⁴ state that verbal encouragement was given for HHD and ID but not stated with MMT. However, the value of verbal encouragement is poorly represented in the literature on SCI.

Measuring strength changes in subjects with SCI can provide evidence of recovery or deterioration of motor output and thereby provide diagnostic information as well as direct the efficiency of rehabilitation interventions.¹⁰ Therefore, clinicians need to review the

evidence to support best practice when selecting a method for assessing muscle strength, not just in adults with SCI but for all diagnostic groups.

Hand Held Dynamometry

A further objective of this systematic review was to determine whether it was appropriate for HHD to become the standard tool for measuring muscle strength for adults with SCI.

This section aims to discuss the result findings specific to HHD. Eight of the 11 included studies researched HHD in relation to participants with SCI.^{4,12,27,28,29,30,31,32} The authors of these studies raised a number of issues relating to the application of HHD for assessing muscle strength in adults with SCI. These include whether the 'make' or 'break' technique is used,¹² the strength of the examiner/ tester,³² whether a fixed lever arm is used,³⁰ isometric contractions versus movement for testing,²⁷ the experience of the examiner,¹² testing procedures.²⁸

A worthy subject of discussion is the type of contraction measured and it is important to note that measurements taken under differing conditions should not be directly compared. Maximum force during eccentric (muscle lengthening) dynamic contractions is greater than that during isometric (static) contractions, which is greater than that during concentric (muscle shortening) contractions.³⁵ When looking at HHD which generally use isometric contractions, then the strength measurement depends on whether a 'make' or 'break' technique is used. Of the eight included studies which used HHD, 1 looked at make or break techniques, 2 looked at make and break techniques, 1 used the break technique and 4 used the make technique. The 1 study which used make or break technique with HHD testing of the postural muscles reported good reliability for intra-tester (ICC > 0.96) and inter-tester (> 0.80).²⁹ However, this is in conflict to other research studies which have demonstrated that the break technique produces higher scores than the make technique,^{12,27} Two of the included studies compared break/make ratios.^{12,27} It has been reported that the break technique results in higher strength values than the make technique,²⁷ which is supported by both of these studies.^{12,27} One explanation for this is that despite make and break contractions both being isometric, the make technique is closer to a concentric contraction and the break technique is more related to an eccentric contraction. Therefore, it should be clearly stated which technique is used when measuring muscle strength and the two techniques are not considered interchangeable.

It is also worth noting that in the study by Herbison et al.³¹ comparing MMT with HHD, the researchers used elbow flexors with a minimum of grade 3.5 on the modified MRC scale. The authors state that the reason for this was because testing elbow flexors with a HHD is not possible for the lower grades of strength unless the extremity is placed in the gravity eliminated position.³¹ The authors chose not to place the limb in a gravity eliminated position in order to maintain consistency.³¹ Therefore, the distinction that should be made is that it is possible to test muscle strength in the lower MMT grades with HHD, just that the authors chose not to so their comments only relate to grades 3.5-5.³¹ However, if a patient presents with no movement (grade 0) or only a flicker of movement (grade 1) it maybe that a dynamometry score is not possible but a notation would be added instead for that part of the assessment.

The included study by Aufsesser et al.³⁰ reported good intra-tester reliability but poor inter-tester reliability with HHD. The authors go on to state that the inter-tester variability found in their study was likely caused by the difference in the tester's strength and the

movement of the tester's arm during data collection.³⁰ The examiners in their study was a university professor with extensive HHD testing experience and an experienced physical therapist but with little HHD experience.³⁰ However, both examiners assisted in the development of the testing protocol which also underwent pilot testing.³⁰ The authors recommend that the HHD, instead of being held by the examiner should be attached to a fixed lever arm that would not move during testing.³⁰ None of the included studies used fixed lever HHD. Fixed dynamometers have been shown to be reliable in other participant populations.³⁵ However, one of the advantages of using HHD is their portability which would be limited once they have to be secured to a fixed point.

Differences were also found in the reliability of ID and HHD when paraplegics were compared with tetraplegics.^{4,28} Adults with SCI may have difficulty assuming certain positions particularly in gravity eliminated positions, therefore, one of the challenges to standardisation of testing position is gravity.¹⁰ However, this could be said of all 3 methods of assessing muscle strength.

Tester strength is another issue which is raised in a number of HHD studies. May et al.²⁸ found that the stronger tester produced higher reliability scores than the weaker tester for the stronger muscle groups. However, this maybe less of an issue with SCI compared to athletic populations due to the fact that with patient populations there is often muscle weakness present. In addition, stability and resistance are more easily provided when persons with SCI are tested compared with competitive athletes, since many of the tests of subjects with SCI are performed seated, allowing the examiner to position themselves to provide adequate resistance.¹⁰ However, it has been reported that the capacity of inexperienced therapists of only average strength to reliably assess muscle strength with HHD increases with experience.³⁵ Of the 8 included studies which examined HHD, 5 used experienced examiners,^{4,27,28,31,32} 1 used inexperienced physiotherapy students¹² and 2 used experienced and inexperienced examiners.^{29,30} Only the 1 study by Aufsesser et al.³⁰ reported poor inter-tester reliability with HHD. However, tester strength can be attenuated by standardisation of procedures and putting the tester at a mechanical advantage to the patient so less strength by the examiner is required.

In the 8 included studies the majority used 3 trials for HHD testing which usually preceded practise/ instrument familiarisation.^{2,30} However, the studies differed on whether they used the mean, best score or a specific numbered score (i.e. 2 practices and then the recorded measurement trial).³⁰ The number of trials is an unresolved issue. Some studies on different participant populations that the first trial was always the highest score and recommend only one trial after a submaximal practice to attenuate risk of injury, whereas others recommend the use of averages particularly in inexperienced examiners until their capacity to obtain reliable single trials is established. The belief being that repeated scores will regress towards the true mean.³⁹ The study by Aufsesser et al.³⁰ found the average scores compared with single trial score produced higher intra-tester and inter-tester reliability in all cases, although they were the only study to compare both. However, as with all other variants, the clinician must record their testing procedure and be consistent from test to test.

Only 1 study used movement rather than an isometric contraction for their research.²⁷ The primary objective of their study was to determine the relationship between strength measurement and angular velocity of which intra-tester reliability was a component of their research.²⁷ HHD are becoming more technologically advanced so that the

capabilities of the devices are being enhanced all of the time. This is an exciting progression because the divide between the highly complex ID and the more portable HHD is narrowing. A similar analogy could be made between the size of desktop computers and tablets. However, just like the computer example there are reasons why you may select one device over another. It is exciting that HHD can incorporate movement and muscle contractions other than the traditional isometric contraction, however, due to their hand held nature HHD may never have the advanced capabilities of ID but does this matter? Only further, directed research into this area could provide answers. Burns and Spaniers'²⁷ research gives good insight into the make/break ratio as varying speeds but it is an isolated study in this systematic review so less weight can be given to their findings.

Standardisation of testing procedures as discussed earlier is important regardless of the method used. Unlike MMT where there is clearly documented testing procedures (such as those stated in AIS manual) HHD may be more diverse. HHD is the name given to a number of devices by different manufacturers. Therefore, standardisation may vary from device to device and condition to condition. Therefore, standardisation should be performed at a local level. Each clinic should standardise their testing procedures, which can be based on information already provided in user manuals and all examiners should be trained in this with clear documentation. Factors which should be standardised and clearly documented are device calibration, patient position per test, anatomical placement of the device on the patient, technique used (make or break), verbal encouragement, blinding of results, number of practise and measurement trials, single or average scores to name a few. Different HHD devices should not be used interchangeably and even 2 devices of the same make may provide differing results therefore, devices in a research and clinical setting should be clearly labelled and used on all subsequent testing where possible just like individuals may weigh differently on different weight measurement scales.

Instrumented strength tests hold definite advantages over traditional MMT and therefore it is not surprising that these instruments have been used over the last 40 years to document the strength of patients with numerous diagnoses.³⁹

Therefore to conclude, the results for the reliability, validity, responsiveness and interpretability of HHD is favourable particularly when compared to MMT and ID in adults with SCI.

Assumptions

In developing the protocol for this systematic review a number of assumptions were made. It was assumed that the assessment of muscle strength was an important component of the initial assessment for a myriad of clinicians and that this would remain important into the future. This is probably a fair assumption to make as muscle strength assessment is contained within the undergraduate training of a number of professions including physiotherapists and doctors. Muscle strength assessment is also a core component of any neurological or musculo-skeletal examination and most private practices and hospital environments list the results of such an exam in any proforma which is developed. The other assumption is that clinicians are able to access and integrate dynamometry, whether it is HHD or ID, to clinical practice. Dynamometry devices are readily available worldwide at varying costs so are easily accessible if the clinician deems appropriate. MMT has been utilised in clinical environment since the early part of the 20th century and dynamometry first emerged some years later. Therefore it was also assumed that most clinicians reading this review would have a basic understanding of these methods. The other assumption made regarding the participant population of SCI was that again most readers would be familiar with the injury and that the prevalence was sufficient to be relevant to future practice and research.

Limitations

There were a number of limitations of this systematic review. The first limitation was that the studies themselves all used convenience sampling which does mean that from a statistical sampling point of view the results of this review can not be generally applied to a larger population. More men were involved as participants in the included studies and in general the heterogeneity of the participants in terms of level of injury, muscles/joints assessed and the method of assessment used also restricts generalisability,. Only 5 studies looked at MMT (including 2 in isolation^{24,25}), 3 looked at ID (including 1 study in isolation²⁶) and 8 studies included HHD^{4,12,27,28,29,30,31,32} (4 in isolation). Therefore, only 4 studies compared muscle assessment methods in their research and only 1 study compared all three^{4,28,31,32}. The heterogeneity also prevented a pooling of the results -a meta-analysis, so results were in narrative and diagrammatic format. Another limitation included the assessment of methodological quality. The COSMIN tool was used as this was deemed 'best fit', however, the tool was originally devised for health related questionnaires so a number of boxes and questions were not applicable. Future systematic reviews may be rewarded with a most specific methodological quality assessment tool currently being developed.

Delimitations

A number of delimitations were also made to define the boundaries of this systematic review. The first being that the literature search only included studies published in English language. This was due to the issues regarding translation, however, considering that only 11 studies were included in the review, there may have been more if studies were not filtered by language. Restriction was also placed on age with only adults considered aged 15 years and older. There may have been studies which included younger participants which may have introduced new findings which were otherwise missed. SCI was chosen as the focus rather than another condition or muscle/ joint. This was done for a number of reasons. Firstly, systematic reviews have already been written on

dynamometry, grip strength and MMT.^{11,13,36} Also when scoping the literature during the protocol stage there were a lot of diverse research papers which varied from the joint selected; whether or not the participants had disease characteristics or were from normal or athletic populations. SCI seemed to yield the most papers which included MMT, ID and or HHD. In addition by studying SCI the range of muscle weakness present appeared to transcend all grades on the MRC scale which would then present specific issues when using dynamometry such as trunk control in tetraplegia or lack thereof. Although the results can not be generalised to other populations it does however, provide discussion points which could be considered by clinicians when integrating these methods into their practice with a broader spectrum of patient groups. It was important for this systematic review to compare the 3 methods of assessing muscle strength because one of the objectives was to determine whether there was enough supportive evidence to recommend the use of HHD as the standard tool for SCI. At the time of writing this (January 2014) it is believed to be the only systematic review comparing the 3 methods regardless of condition. It was felt that more useful discussion could be made by comparing the methods rather than considering them in isolation. For example as previously discussed MMT as part of the ASIA standards has been shown to be reliable and maybe considered sufficiently valid for diagnostic purposes but MMT is not as sensitive to change as dynamometry therefore, MMT has its own limitations when used as an outcome measure or detailed charting of muscle strength/ weakness over time, pertinent to rehabilitation or prognosis. ID is often reported as unwieldy and expensive.¹⁰ despite also being described as the gold standard,¹¹ whereas MMT is considered the standard tool. Therefore by comparing these 3 methods and reviewing the advantages and issues related to each one, it was felt that a clearer understanding in the assessment of muscle strength in SCI would emerge.

Chapter 5: Conclusions

As part of any assessment process, clinicians will conduct subjective and objective examinations. The information that they derive from this will form the basis of any diagnosis, prognosis and evaluation. The measurement of functioning and disability can take many forms as there are numerous instruments available to the clinician. However, there are a number of issues related to the development and application of measurement instruments. These include the selection of an instrument out of a wide range available, using an instrument in diverse diagnostic groups, and the use of instruments in clinical practice.⁴⁰ This chapter aims to address these issues and conclude the results from this systematic review.

Implications for practice

Clinicians are often under increasing pressure from their work setting, outside agencies, colleagues and patients to introduce measurement in clinical practice especially to evaluate rehabilitation outcomes as part of best practice.⁴⁰ However, it is important that time for measurement is not taken away from patient treatment and that measurement is not simply taken for measurement sake but is directly advantageous to the patient. In addition, the measurement instrument used should have evidence to support its reliability, validity, responsiveness and interpretability otherwise the results and relevance can be undermined. With regard to the measurement of muscle strength, which is an important component of the objective examination, clinicians should be aware of the clinimetrics of the methods available. However, method selection will not just be based on research but also on instrument availability, ease of use, knowledge of use, time taken to set up, clinical relevance of data and cost.

MMT has the advantage in most of these areas as it requires no equipment, is taught at undergraduate level and requires very little time to set up and complete. However, this systematic review has also shown that MMT is not very responsive to changes in muscle strength over time and that changes shown with dynamometry testing did not always translate to changes in MMT grade. Significant overlapping of dynamometry scores were seen at the higher MMT grades all of which affects the overall validity of the method. Good intra-tester reliability was demonstrated but questionable inter-tester. ID on the other hand is expensive, not always available and time consuming to set up. But the advantages are the sophistication of the device and its capabilities to test all types of muscle contraction. However, despite ID being labelled as the gold standard there were issues seen with ID particularly in relation to tetraplegia.

HHD has many of the advantages of MMT and ID; it uses ratio data and is responsive to change in strength measures similar to ID. It is portable and easy to use in a bed side environment like MMT. However, it is much cheaper than ID but more expensive than MMT so clinicians would need to be committed and motivated in purchasing a device and educating their staff on its usage. Some of the disadvantages to HHD are not inherent to the device but apply to all of the methods for measuring muscle strength such as standardisation of procedure, examiner education and accurate documentation. However, HHD has literature to support its reliability and validity as a measurement tool which may support its wider use in clinical practice.

This is an increasingly technological age. Patients and families are better informed with greater access to medical information than ever before. Clinicians are encouraged to embrace the research and use the technological advancements available to them as long as they are relevant to the patient. Does this systematic review provide sufficient evidence to change clinical practice? This review presents the research currently available for the measurement of muscle strength in SCI. It suggests that clinimetrically HHD has the more supportive evidence in its use but it also had the most research studies to draw knowledge from. MMT as part of the AIS may still be the most appropriate method and if more sophisticated measurement is required then ID may be more appropriate. But given the evidence to support the reliability (both inter and intra tester) by experienced and inexperienced examiners, the validity, responsiveness to change, ease of interpreting the results, it certainly implies that there are significant advantages to the extended use of HHD in clinical practice. HHD is able to provide the clinician and patient with reliable measurement of muscle strength during examination which can be retested over time to demonstrate change. HHD uses kilogram or Newton measurement data which is often easier for the patient to understand rather than a grade of movement. This can assist with the treatment selection process and motivation. It can provide outcome measures for evaluation and assist with prognostic markers. Whichever method the clinician uses to assess muscle strength, it is hoped that this review has assisted in the evidence of best practice for patients with SCI.

Implications for research

There has been a lot of research in the area of MMT, HHD, ID in isolation over a variety of conditions, joints, normal and athletic populations. However, although a number of the studies are drawing similar conclusions, due to the specificity of the participant population, study design and type of method/device generalisability is problematic. Further studies are required in the area of SCI and in particular researchers need to address the issue of heterogeneity and consider the level of credibility of their studies. This includes the choice of method selected, whether randomised controlled trials would be appropriate, as well as controlling the variability of sampling techniques, sample size, outcomes measures and statistical tests used.

An area already discussed is related to type of movement assessed which can be further developed to include function. The fourth point made by Bohannon³⁵ documented earlier in Chapter 4 was that strength measurements should be predictive of other variables of potentially greater importance. Typically the movement assessed is an open chained movement, which means that the distal segment is free to move, such as a biceps curl where the hand is free to move. Functional movement is more typically closed chain such as the squatting motion in sit to stand. Therefore, exploration through research is required on the relationship between muscle strength assessment and functional outcomes, or even as a predictor of prognosis. There is research on the relationship of grip strength and mortality in elderly participants³⁴ A relationship between muscle strength and wheelchair propulsion ability in paraplegic men has been identified in one study.⁴¹ Strength tests with dynamometry may also assist in injury prevention. Souza et al.⁴² compared upper limb strength of 28 persons with paraplegia who were manual wheelchair users, with non-disabled control subjects using ID. They found that those SCI subjects with high level of injuries had lower shoulder extension strength measurements than the non-disabled

controls. Whereas the SCI with low injury levels exhibited greater shoulder abduction strength compared with the control group⁴²

Reported traumatic SCI annual incidence rates worldwide yielded values ranging from 12.1 to 57.8 cases per million, with motor vehicle collisions, falls, violence and sports activities being the leading causes.³⁴ Causes of non-traumatic SCI include spinal stenosis, spinal tumour compression, congenital deformities and inflammatory conditions. The aetiology of the SCI is interesting. Van der Berg et al.³⁴ in their systematic review on incidence of SCI worldwide found that the incidence rate of non-traumatic SCI increased steadily with age as might be expected with a number of the causes being related to the aging process. However, with traumatic SCI most of the studies they identified described a bimodal distribution with the first peak seen in young adults between 15 to 29 years and the second peak in the age ranges of 40-49, >50 years and >65 years, with falls being the primary cause in 57 % of patients older than 60 years and 89% of those older than 80 years.³⁴ Furthermore, a number of studies have found a relationship between weak knee and ankle muscles and a higher incidence of falling, as well as a relationship between poor grip strength and falls.^{43,44,45} Therefore this is another exciting area to direct future research in not just the prevention of SCI through falls, but also looking at falling rates amongst the SCI population and further prevention of this.

Further research is recommended specific to HHD related to feasibility, clinician and patient satisfaction, ease of use, cost analysis and clinimetric studies over more joints/muscle groups before it can be recommended as the standard tool.

General Conclusion

The objective of this systematic review was to look at the best clinimetric evidence available on the assessment of muscle strength in SCI using MMT, ID and HHD, and whether there was support in the literature to recommend HHD as the standard assessment tool.

Eleven studies were included with the results presented in diagrammatic and narrative form due to their heterogeneity.

Despite MMT being around for almost 100 years and dynamometry's emergence in the last 45-55 years, all of the included studies were written in the last 21 years. Therefore, this is an exciting area of research as clinician's start to consider clinimetric studies and question the reliability, validity, responsiveness and interpretability of the assessment tools utilised. Muscle strength is an important component of the objective examination of a patient and the increasing prevalence of SCI means that this could be considered an area of interest to many.

The results of this systematic review have lead to a number of discussion points and conclusions.

Firstly, from the 2 studies which included MMT as part of the ASIA examination, questions were raised as to the inter-tester reliability of MMT and whether differences in MRC scale grades given to an individual patient would change the overall ASIA score assigned to that patient.^{24,25} However, as individual MMT grades are combined to give an

overall score for the ASIA examination, this may pose less of an issue. Overall, MMT was shown to have varied inter-tester reliability, although scores were improved following standardisation of the procedure.

However, when MMT was compared to another method of measuring muscle strength and studied away from the context of the ASIA examination, this gave a different picture. A number of the included studies questioned the reliability in a number of the MRC grades, with poor reliability shown at grades 2, 3, 4²⁵ and grades 4 or above.^{4,32} Despite a positive correlation of increasing scores for MMT and dynamometry; significant overlapping of dynamometry scores in the higher MMT grades were observed which questions the validity of MMT.^{4,24,31,32} Universally the sensitivity of MMT was questioned, given that it uses ordinal data. Therefore, in light of these findings, the interpretability of MMT scores can also be questioned, particularly at the higher grades; a patient classified as normal on the MRC scale but still exhibiting muscle strength improvements with HHD can not be classified as more normal on the MRC scale.^{4,31}

ID has been considered the gold standard of muscle strength assessment tools and certainly it has its advantages over the other methods due to the types of muscle contraction which can be assessed and the amount of data which can be collated.¹¹ However, 2 of the included studies did question the reliability of ID in relation to tetraplegics due to the standardised positioning of these patients particularly with elbow movement, given the differing trunk stability of these patients in comparison to other patient groups.^{4,28} However, with MMT and ID the results must be considered in the context of the small numbers of studies included.

HHD yielded the most included studies. Five of the 6 studies which considered reliability concluded that HHD was intra-tester reliable,^{12,27,28,29,30} 2 supported inter-tester reliability^{27,29} and only 1 study concluded that HHD had poor inter-tester reliability.³⁰ Reliability was also demonstrated in one study which looked at HHD with movement rather than an isometric contraction.²⁷ HHD also performed well in the 3 studies which looked at its validity.^{4,31,32} HHD was responsive to change in scores in the 3 included studies and could be considered easy to interpret due to the ratio data, often measured in kilograms or newtons.^{4,28,31,32}

However, despite these conclusions, it must be highlighted that the dynamometry studies only researched specific muscle groups in the upper limb and trunk and no lower limb muscles were assessed. Whereas the MMT included equal number of upper limb and lower limb muscle groups as defined by the cervical and lumbar neurological spinal levels.

The results also need to be considered within the context of the heterogeneity of the patient population, including age and gender and that convenience rather than random sampling was utilised in all studies. Although all of the patients had SCI they differed in the severity of the injury and therefore the resulting disability (paraplegia versus tetraplegia and ASIA grade A-D). More men were represented in the studies with a wide range of ages and time since injury.

To conclude, based on the findings of this systematic review, when considering the clinimetrics of the 3 methods for assessing muscle strength in SCI there does appear to be support in the literature to recommend the extended use of HHD as a reliable, valid,

responsive, interpretable tool adjunctive to patient assessment and treatment for adults with SCI.

References

1. ASIA impairment scale. URL:
http://sci.rutgers.edu/dynarticles/2001_Classif_worksheet.pdf
2. O'Connor, PJ. Prevalence of spinal cord injury in Australia. *Spinal Cord*. 2005; 43: 42-46.
3. Wyss JF, Patel A. Therapeutic programming for musculoskeletal disorders. Demos Medical Publishing. 2013; 1: 24-25.
4. Noreau, L, Vachon, J. Comparison of three methods to assess muscular strength in individuals with spinal cord injury. *Spinal Cord*. 1998; 36: 716-723.
5. Galea, M. Introducing clinimetrics. *Aust J Physiother*. 2005; 5: 139-140.
6. Marino RJ, Jones L, Kirshblum S, Tal J, Dasgupta A. Reliability and repeatability of the motor and sensory examination of the international standards for neurological classification of spinal cord injury. *J spinal cord med*. 2008; 31(): 166-170.
7. Medical Research Council. Aids to the examination of the peripheral nervous system. Memorandum no.45, Her Majesty's stationary office. London. 1981.
8. Bohannon, R., Andrews, AW. Interrater reliability of hand held dynamometry. *Phys Ther*. 1987; 67: 931-933.
9. Kellin, B, McKeon, PO, Gontkof, LM. Hand-held dynamometry: Reliability of lower extremity muscle testing in healthy, physically active, young adults. *J Sport Rehabil*. 2008; 17: 160-170.
10. Sisto, SA, Dyson-Hudson, T. Dynamometry testing in spinal cord injury. *J Rehabil Res Dev*. 2007; 44: 123-136.
11. Stark, T, Walker, B, Phillips, J, Fejer, R, Beck, R. Hand held dynamometry correlation with the gold standard isokinetic dynamometry: a systematic review. *Am J Phys Med Rehabil*. 2011; 3: 472-479.
12. Burns SP, Breuninger A, Kaplan C, Marin H. Break-technique handheld dynamometry: relation between angular velocity and strength measurements. *Arch Phys Med Rehabil*. 2005; 84(1): 22-29.
13. Mafi P, Mafi R, Khan W. A systematic review of dynamometry and its role in hand trauma assessment. *Open Orthop J*. 2012; 6: 95-102.
14. Mokkink, LB, Terwee, CB, Patrick, DL, Alonso, J, Stratford, PW, Knol, DJ. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. *J Clin Epidemiol*. 2010;63:737-745.
15. Portney, LG, Watkins MP. Foundations of clinical research. Pearson Prentice Hall.2009; 3rd Ed .
16. Sim J, Arnell, P.. Measurement Validity in Physical Therapy Research. *Phys Ther*.1993; 73: 102-110.
17. Tufanaru C, Huang WJ, Tsay S-F, Chou S-S. Statistics for systematic reviews.

Lippincott-Joanna Biggs Institute.2012; 11.

18. Mokkink, LB, Terwee, CB, Patrick, DL, Alonso, J, Stratford, PW, Knol, DJ. COSMIN checklist manual. 2010: URL http://cosmin.nl/the-cosmin-checklist_8_5.html

19. Mokkink, LB, Terwee, CB, Patrick, DL, Alonso, J, Stratford, PW, Knol, DJ. The COSMIN checklist for assessing the methodological quality of studies: an International Delphi study. Qual Life Res. 2010;

20. Dobson F, Choi YM, Hall M, Hinman RS. Clinimetric properties of observer assessed impairment tests used to evaluate hip and groin impairments: A systematic review. Arth Care Res, 2012;64 (10):1565-1575.

21. Peek K, Pensuk, P. Muscle strength in adults with spinal cord injury: A systematic review protocol of manual muscle testing, isokinetic and hand held dynamometry clinimetrics. JBI database of systematic reviews and implementation reports. 2013;9(11):105-119.

22. Cohen ME, Ditunno JF, Donovan WH, Maynard FM. A test of the 1992 international standards for neurological and functional classification of spinal cord injury. Spinal cord. 1998; 36: 554-560.

23. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW.. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010; 10(22): 8-16.

24. Jonsson M, Tollback A, Gonzales H, Borg J. Inter-rater reliability of the 1992 international standards for neurological and functional classification of incomplete spinal cord injury. Spinal cord. 2000 Nov;38(11):675-9.

25. Savic G, Bergstrom EM, Frankel HL, Jamous MA, Jones PW. Inter-rater reliability of motor and sensory examinations performed according to American Spinal Injury Association standards. Spinal cord. 2007 Jun;45(6):444-51.

26. Kakebeeke TH, Lechner HE, Handshin C.. Reproducibility analysis of isokinetic strength measurements of shoulder and elbow muscles in subjects with spinal cord injury. Isokinet Exerc Sci. 2005; 13: 279-284.

27. Burns SP, Spanier DE. Break-technique handheld dynamometry: relation between angular velocity and strength measurements. Arch Phys Med Rehabil. .2005; 86(7): 1420-6.

28. May LA, Burnham RS, Steadward RD. Assessment of isokinetic and hand-held dynamometer measures of shoulder rotator strength among individuals with spinal cord injury. Arch Phys Med Rehabil.1997; 78: 251-255.

29. Larson CA, Tezak WD, Shepherd Malley S, Thornton W.. Assessment of postural muscle strength in sitting: reliability of measures obtained with hand held dynamometry. J Neurol Phys Ther : .2010; 34: 24-31.

30. Aufsesser PM, Horvat M, Austin R.. The reliability of hand held muscle testers with individuals with spinal cord injury. Clin Kinesiol.2003; 57(4): 71-75.

31. Herbison GJ, Isaac Z, Cohen ME, Ditunno JF, Jr. Strength post-spinal cord injury: myometer vs manual muscle test. *Spinal cord*.1996; 34: 543-548.
32. Schwartz S, Cohen ME, Herbison GJ, Shah A. Relationship between two measures of upper extremity strength: manual muscle test compared to hand-held myometry. *Arch Phys Med Rehabil* .1992; 73(11):1063-1068.
33. Sim, J. Wright, CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys Ther*. 2005. 85 (3): 257-267.
34. van den Berg MEL, Castellote JM, Mahillo-Fernandez I, de Pedro-Cuesta J. Incidence of spinal cord injury worldwide: A systematic review. *Neuroepidemiol*. 2010; 34: 184-192.
35. Bohannon RW. The clinical measurement of strength. *Clin Rehabil*.1987; 1: 5-16.
36. Cuthbert SC, Goodheart GJ. On the reliability and validity of manual muscle testing: a literature review. *Chiropr Osteopat*. 2007; 15(4): 23-46.
37. de Vet HCW, Terwee CB, Bouter LM. Current challenges in clinimetrics. *J Clin Epidemiol*.2003; 56(12): 1137-1141.
38. Wadsworth CT, Krishnan R, Sear M, Harrold J, Nielsen DH.. Intrarater reliability of manual muscle testing and hand held dynamometric muscle testing. *Phys Ther*.1987; 67(9): 1342-1347.
39. Bohannon RW. Hand held dynamometry: factors influencing reliability and validity. *Clin Rehabil*. 1997;11:263-264.
40. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments. *J Rehabil Med*. 2005; 37: 193-201.
41. Armstrong LE, Winant DM, Swasey PR. Using isokinetic dynamometry to test ambulatory patients with multiple sclerosis. *Phys Ther*.1986; 63: 1274-1279.
42. Souza AL, Boninger ML, Fitzgerald SG, Shimada SD, Cooper, RA, Ambrosio F. Upper limb strength in individuals with spinal cord injury who use manual wheelchairs. *J Spinal Cord Med*. 2005; 28(1): 26-32.
43. Takazawa K, Arisawa K, Honda S, Shibata Y, Saito, H. Lower extremity muscle forces measured by hand held dynamometer and the risk of falls among day care users in Japan. *Disabil Rehabil*.2003; 25: 399-404.
44. Whipple RH, Wolfson LI, Amerman PM. The relationship of knee and ankle weakness to falls in nursing home residents: an isokinetic study. *J Am Geriatr Soc*.1987; 35(1): 13-20.
45. Stalenhoef PA, Diederiks JPM, Knottnerus JA, Kester, ADM, Crebolder HFJM. A risk model for the prediction of recurrent falls in community dwelling elderly: A prospective cohort study. *journal of clinical epidemiology*.2002; 55(11): 1088-1094.

Appendix I: Data Extraction Instrument

		Study
Authors, year and title		
Muscles assessed in study		
Types of participants including diagnosis specifics	Age	
	Gender	
	No. of participants	
	Diagnosis and co-morbidities	
	ASIA score (if known)	
	Time since injury	
Method of muscle strength test used	Manual Muscle Test	
	Isokinetic Dynamometry	
	Hand Held Dynamometry	
Trained examiner (yrs of experience)		
Blinded examiner?		
Clinimetric domain evaluated	Reliability	
	Validity	
	Responsiveness	
	Interpretability	
Statistical analysis used		
Results		
Authors Conclusions		
Reviewers Comments		
COSMIN score		

Appendix II: Critical Appraisal Instrument

COSMIN checklist with 4-point scale¹⁸

Instructions

This version of the COSMIN checklist is recommended for use in systematic reviews of measurement properties. With this version it is possible to calculate overall methodological quality scores per study on a measurement property. A methodological quality score per box is obtained by taking the lowest rating of any item in a box ('worse score counts'). For example, if for a reliability study one item in the box 'Reliability' is scored poor, the methodological quality of that reliability study is rated as poor. The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g. norm scores, floor-ceiling effects, minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore no scoring system was developed for these boxes.

Step 1. Evaluated measurement properties in the article

(The COSMIN checklist is a modular tool which means that it is not necessary to complete the whole checklist when evaluating the quality of a particular study. The following includes the COSMIN boxes to be used.)

Reliability

Box B

Criterion validity

Box H

Responsiveness

Box I

Interpretability

Box J

Step 2. Determining if a study meets the standards for good methodological quality

Box A. Reliability:

Reliability	Excellent	Good	Fair	Poor
Design requirements 1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the		Other minor methodological flaws in the design	Other important methodological flaws in the

	design or execution of the study		or execution of the study	design or execution of the study
Statistical methods 11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12 for dichotomous/nominal/ordi nal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13 for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14 for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box H. Criterion validity

	1	2	3	4
Design requirements 1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (< 30)
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box I. Responsiveness

	1	2	3	4
Design requirements	Percentage of missing items described	Percentage of missing items NOT described		
1 Was the percentage of missing items given?				
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥ 100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (< 30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No Longitudinal design used
5 Was the time interval stated?	Time interval stated			Time interval NOT stated
6 If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7 Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what	Unclear what was expected

8 Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?			was expected	
9 Was the expected direction of correlations or mean differences of the change scores of the instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10 Were the expected absolute or relative magnitude of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
11 Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12 Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13 Were there any	No other		Other minor	Other

important flaws in the design or methods of the study?	important methodological flaws in the design or execution of the study		methodological flaws in the design or execution of the study	important methodological flaws in the design or execution of the study
Statistical methods 14 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate
Design requirement for comparison to a gold standard For constructs for which a gold standard was available: 15 Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box J. Interpretability

	1	2	3	4
1 Was the percentage of missing items given?				

2 Was there a description of how missing items were handled?				
3 Was the sample size included in the analysis adequate?				
4 Was the distribution of the (total) scores in the study sample described?				
5 Was the percentage of the respondents who had the lowest possible (total) score described?				
6 Was the percentage of the respondents who had the highest possible (total) score described?				
7 Were scores and change scores (i.e. means and SD)				

presented for relevant (sub) groups? e.g. for normative groups, subgroups of patients, or the general population				
8 Was the minimal important change (MIC) or the minimal important difference (MID) determined?				
9 Were there any important flaws in the design or methods of the study?				

The COSMIN checklist results:

Key:

E: Excellent

G: Good

F: Fair

P: Poor

?: not stated

N/A: not applicable

The study number related to studies as listed in the data extraction instrument in Appendix II and included studies Appendix IV

Box B. Reliability: relative measures (including test-retest reliability, inter-tester reliability and intra-tester reliability)

	Study number										
	1	2	3	4	5	6	7	8	9	10	11
Design requirements	N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A		N/A
1 Was the percentage of missing items given?											
2 Was there a description of how missing items were handled?	N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A		N/A
3 Was the sample size included in the analysis adequate?	P	P	G	P		P	P	E	P		P
4 Were at least two measurements available?	E	E	E	E		E	E	E	E		E
5 Were the administrations independent?	E	G	E	E		E	E	E	E		G
6 Was the time interval stated?	E	E	E	E		E	E	E	F		G
7 Were patients stable in the interim period on the construct to be measured?	G	G	G	G		G	G	E	F		G
8 Was the time interval appropriate?	F	E	E	E		E	E	E	E		E
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	E	E	E	E		E	E	E	G		E
10 Were there any important flaws in the design or methods of the study?	F	E	E	F		E	E	E	F		E
Statistical methods	E	E	E	E		N/A	N/A	G	E		G
11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?											
12 for dichotomous/nominal/ordinal scores: Was kappa calculated?	N/A	N/A	N/A	N/A		E	E	N/A	N/A		P
13 for ordinal scores: Was a weighted kappa calculated?	N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A		N/A
14 for ordinal scores: Was the weighting scheme described? e.g.	N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A		N/A

linear, quadratic											
Overall COSMIN score and question/s number the score relates to ().	P (1)	P (1)	G (3,7)	P (1)		P (1)	P (1)	G (11)	P (1)		P (1, 12)

Box H. Criterion validity

	Study number										
	1	2	3	4	5	6	7	8	9	10	11
Design requirements			N/A		N/A			N/A		N/A	N/A
1 Was the percentage of missing items given?											
2 Was there a description of how missing items were handled?			N/A		N/A			N/A		N/A	N/A
3 Was the sample size included in the analysis adequate?			G		F			E		G	
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?			E		G			P		P	
5 Were there any important flaws in the design or methods of the study?			E		G			G		F	
Overall COSMIN score and question/s number the score relates to ().			G (3)					P (4)		P (4)	

Box I. Responsiveness

	Study number										
	1	2	3	4	5	6	7	8	9	10	11
Design requirements					N/A			N/A		N/A	
1 Was the percentage of missing items given?											
2 Was there a description of how missing items were handled?					N/A			N/A		N/A	
3 Was the sample size included in the analysis adequate?					E			E		G	
4 Was a longitudinal design with at least two measurement used?					E			E		E	
5 Was the time interval stated?					E			E		E	
6 If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?					G			G		G	
7 Was a proportion of the patients changed (i.e. improvement or deterioration)?					N/A			N/A		N/A	
For constructs for which a gold standard was not available:					N/A			N/A		N/A	
8 Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?											
9 Was the expected direction of correlations or mean differences of the change					N/A			N/A		N/A	

scores of the instruments included in these hypotheses?											
10 Were the expected absolute or relative magnitude of correlations or mean differences of the change scores of the instruments included in these hypotheses?					N/A			N/A		N/A	
11 Was an adequate description provided of the comparator instrument(s)?					N/A			E		E	
12 Were the measurement properties of the comparator instrument(s) adequately described?					N/A			E		E	
13 Were there any important flaws in the design or methods of the study?					N/A			E		E	
Statistical methods 14 Were design and statistical methods adequate for the hypotheses to be tested?					N/A			N/A		N/A	
Design requirement for comparison to a gold standard For constructs for which a gold standard was available: 15 Can the criterion for change be considered as a reasonable gold standard?					E			N/A		N/A	
16 Were there any important flaws in the					F			N/A		N/A	

design or methods of the study?											
Overall COSMIN score and question/s number the score relates to ().					F (16)			G (6)		G (3,6)	

Box J. Interpretability

Study number											
	1	2	3	4	5	6	7	8	9	10	11
1 Was the percentage of missing items given?					N/A			N/A			
2 Was there a description of how missing items were handled?					N/A			N/A			
3 Was the sample size included in the analysis adequate?					yes			yes			
4 Was the distribution of the (total) scores in the study sample described?					N/A			N/A			
5 Was the percentage of the respondents who had the lowest possible (total) score described?					N/A			N/A			

6 Was the percentage of the respondents who had the highest possible (total) score described?					N/A			N/A			
7 Were scores and change scores (i.e. means and SD) presented for relevant (sub) groups? e.g. for normative groups, subgroups of patients, or the general population					N/A			N/A			
8 Was the minimal important change (MIC) or the minimal important difference (MID) determined?					N/A			N/A			
9 Were there any important flaws in the design or methods of the study?					no			no			

Appendix III: Detailed search strategy

Database	Key terms used for search and number of papers found		
	Spinal cord injury AND manual muscle test	Spinal cord injury AND dynamometer/ dynamometry/ myometer/myometry	Spinal cord injury AND manual muscle test AND dynamometer/ dynamometry/ myometer/myometry
Pubmed	79	57	3
CINAHL	42	57	6
OT seeker	0	0	0
Cochrane	5	1	0
SportDISCUS	48	89	2
PEDro	2	1	0
Australian Physiotherapy Journal	2	1	0
UK Physiotherapy Journal	112	3	1
Spinal Cord	369	54	30
Mednar	1,060	96	1
Total before excluding duplicates	1,719	359	43
Total for all search terms before excluding duplicates: 2,121			

Limitations for all searches included English language, adults,

The Electronic Databases searched from inception to 2013 were:

- Cochrane database
- PubMed,
- CINAHL,
- SPORTSDiscus,
- OT Seeker,
- PEDro,
- Physiotherapy journals UK and Australia.
- Journal: Spinal Cord
- Mednar database

Appendix IV: Included studies

1. Burns SP, Breuninger A, Kaplan C, Marin H. Break-technique handheld dynamometry: relation between angular velocity and strength measurements. *Arch Phys Med Rehabil.* 2005; 84(1): 22-29.
2. Larson CA, Tezak WD, Shepherd Malley S, Thornton W... Assessment of postural muscle strength in sitting: reliability of measures obtained with hand held dynamometry. *J Neurol Phys Ther.* .2010; 34: 24-31.
3. May LA, Burnham RS, Steadward RD. Assessment of isokinetic and hand-held dynamometer measures of shoulder rotator strength among individuals with spinal cord injury. *Arch Phys Med Rehabil.*1997; 78: 251-255.
4. Kakebeeke TH, Lechner HE, Handshin C. Reproducibility analysis of isokinetic strength measurements of shoulder and elbow muscles in subjects with spinal cord injury. *Isokinet Exerc Sci.*2005; 13: 279-284.
5. Noreau, L, Vachon, J. Comparison of three methods to assess muscular strength in individuals with spinal cord injury. *Spinal Cord.*1998; 36: 716-723.
6. Savic, G Bergstrom EMK, Frankel HL, Jamous MA, Jones PW. Inter-rater reliability of motor and sensory examinations performed according to American Spinal Injury Association standards. *Spinal cord.*2007; 45(): 444-451.
7. Jonsson M, Tollback A, Gonzales H, Borg J. Inter-rater reliability of the 1992 international standards for neurological and functional classification of incomplete spinal cord injury. *Spinal cord.*2000; 38: 675-679.
8. Schwartz S, Cohen ME, Herbison GJ, Shah A. . . Relationship between two measures of upper extremity strength: manual muscle test compared to hand-held myometry. *Arch Phys Med Rehabil.*1992; 73(11): 1063-1068.
9. Aufsesser PM, Horvat M, Austin R... The reliability of hand held muscle testers with individuals with spinal cord injury. *Clin Kinesiol.*2003; 57(4): 71-75.
10. Herbison GJ, Isaac Z, Cohen ME, Ditunno JF, Jr. Strength post-spinal cord injury: myometer vs manual muscle test. *Spinal cord.*1996; 34: 543-548.
11. Burns SP, Spanier DE. Break-technique handheld dynamometry: relation between angular velocity and strength measurements. *Arch Phys Med Rehabil.* .2005; 86(7): 1420-6.

		Study 1
Authors, year		Burns, SP, Breuninger, A, Kaplan, C, Marin H. 2005 Hand held dynamometry in persons with tetraplegia
Muscles assessed in study		elbow flexors, elbow extensors
Types of participants including diagnosis specifics	Age (years)	53.5 +/- 11.7
	Gender	men
	No. of participants	19
	Diagnosis and co-morbidities (no. of participants)	tetraplegia, C4 (1), C5 (12), C6 (6)
	ASIA score (no. of participants)	A (6), B (3), D (10)
	Time since injury	3 < 6 months post injury, 16 > 1 yr post injury
Method of muscle strength test used	Manual Muscle Test	
	Isokinetic Dynamometry	
	Hand Held Dynamometry	Comparison of make v break Chatillon CSD-200 HHD
Trained examiner (yrs of experience)		2 physical therapy students (inexperienced)
Blinded examiner?		no
Clinimetric domain evaluated	Reliability	inter-tester intra-tester
	Validity	
	Responsiveness	
	Interpretability	
Statistical analysis used		intraclass co-efficient (ICC)

Results		ICC	CI 95%
		inter-make (1) 0.91	0.79-0.97
		inter- make (2) 0.94	0.86-0.98
		inter- break (1) 0.93	0.82-0.97
		inter- break (2) 0.94	0.86- 0.98
		intra- make (1) 0.94	0.86-0.98
		intra- make (2) 0.97	0.93-0.98
		intra- break (1) 0.95	0.87-0.98
		intra-break (2) 0.94	0.86-0.98
		B/M ratio 1.38 -1.49	
Authors Conclusions		make and break techniques show high inter- and intra- tester reliability when performed by inexperienced examiners	
Reviewers Comments		Good study issues: no blinding, non-randomised, both examiners present for each test, elbow flex/ext only	
COSMIN score		Poor: sample size, Fair: time interval	

		Study 2																														
Authors, year		Larson, CA, Tezak, WD, Malley, MS. 2010 Assessment of postural muscle strength in sitting: reliability of measures obtained with hand held dynamometry in individuals with spinal cord injury.																														
Muscles assessed in study		postural muscles: anterior, posterior, left lateral, right lateral																														
Types of participants including diagnosis specifics	Age (years)	32.4 +/- 11																														
	Gender	24 men 5 women																														
	No. of participants	29																														
	Diagnosis and co-morbidities (no. of participants)	19 tetraplegics 10 paraplegics																														
	ASIA score (no. of participants)	A (12), B (10), C (5), D (2)																														
	Time since injury	4.5 years +/- 4.6 (0.5 - 19 years)																														
Method of muscle strength test used	Manual Muscle Test																															
	Isokinetic Dynamometry																															
	Hand Held Dynamometry	make or break microFet 2 HHD																														
Trained examiner (yrs of experience)		4 examiners 2 experienced (5 and 25 years) 2 2nd year physio students																														
Blinded examiner?		yes																														
Clinimetric domain evaluated	Reliability	inter-tester intra-tester																														
	Validity																															
	Responsiveness																															
	Interpretability																															
Statistical analysis used		ICC COV																														
Results		<table> <tr> <td></td><td>ICC (unsupported)</td><td>ICC</td></tr> <tr> <td>(supported)</td><td></td><td></td></tr> <tr> <td>anterior</td><td>.98</td><td>.98</td></tr> <tr> <td>posterior</td><td>.99</td><td>.98</td></tr> <tr> <td>right lateral</td><td>.97</td><td>.96</td></tr> <tr> <td>left lateral</td><td>.97</td><td>.97</td></tr> <tr> <td>intra- (supported)</td><td>(unsupported)</td><td></td></tr> <tr> <td>anterior</td><td>.87-.98</td><td>.87-.99</td></tr> <tr> <td>posterior</td><td>.92-.97</td><td>.83-.95</td></tr> <tr> <td>right lateral</td><td>.86-.97</td><td>.85-.95</td></tr> </table>		ICC (unsupported)	ICC	(supported)			anterior	.98	.98	posterior	.99	.98	right lateral	.97	.96	left lateral	.97	.97	intra- (supported)	(unsupported)		anterior	.87-.98	.87-.99	posterior	.92-.97	.83-.95	right lateral	.86-.97	.85-.95
	ICC (unsupported)	ICC																														
(supported)																																
anterior	.98	.98																														
posterior	.99	.98																														
right lateral	.97	.96																														
left lateral	.97	.97																														
intra- (supported)	(unsupported)																															
anterior	.87-.98	.87-.99																														
posterior	.92-.97	.83-.95																														
right lateral	.86-.97	.85-.95																														

		left lateral .84-.92 .85-.95
Authors Conclusions		HHD can be used among different examiners to objectively quantify postural muscle strength in sitting for individuals with SCI.
Reviewers Comments		Generally slightly better results for experienced v student (intra-)
COSMIN score		Poor: sample size Good: assumable pts were stable and measurements independent

		Study 3
Authors, year, title		May, LA, Burnham, RS, Steadward, RD. 1997 Assessment of isokinetic and hand held dynamometer measures of shoulder rotator strength among individuals with spinal cord injury
Muscles assessed in study		shoulder internal and external rotation
Types of participants including diagnosis specifics	Age (years)	26.6 +/- 6.5 (18-42)
	Gender	21 men 4 women
	No. of participants	25, (increasing to 50 with combined results left and right arm)
	Diagnosis and co-morbidities	12 tetraplegics 13 paraplegics
	ASIA score (if known)	
	Time since injury (years)	8.1 +/- 6.9
Method of muscle strength test used	Manual Muscle Test	
	Isokinetic Dynamometry	Cybex 340
	Hand Held Dynamometry	break technique Penny and Giles transducer myometer
Trained examiner (yrs of experience)		yes
Blinded examiner?		yes participants and examiners
Clinimetric domain evaluated	Reliability	intra-tester reliability of HHD and relationship to ID
	Validity	
	Responsiveness	
	Interpretability	
Statistical analysis used		ICC Pearson product moment correlation for relationship (HHD and ID)- comparative
Results		Sh mvt (intra-HHD) ICC (95%CI) comparative

		<p>ext rotn all .94 (.91-.96) .86</p> <p>int rotn all .96 (.94-.98) .88</p> <p>ext para .89 (.80-.94) .83</p> <p>ext tetra .93 (.86-.96) .56</p> <p>int para .92 (.86-.96) .74</p> <p>int tetra .89 (.81-.94) .52</p> <p>Good relationship between HHD and ID co-efficients lower for tetraplegics than paraplegics better co-efficients for stronger examiner than weaker examiner</p>
Auth ors Conclusions		HHD measurement of shoulder rotation in SCI produced reliable results for paraplegics and tetraplegics and were comparable to ID testing for paraplegics. The variability of the isokinetic scores in tetraplegia indicates that this may be an issue with the method of testing (?modified testing protocol)
Reviewers Comments		good study
COSMIN score		Good: sample size (once data combined-50), assumable pts were stable

		study 4
Authors, year, title		Kakebeeke, TH, Lechner, HE, Handschin, C. 2005 Reproducibility analysis of isokinetic strength measurements of shoulder and elbow muscles in subjects with spinal cord injury
Muscles assessed in study		elbow flexion/ extension shoulder in/ext rotation shoulder abd/add
Types of participants including diagnosis specifics	Age (years)	28-59 years
	Gender	?
	No. of participants	10
	Diagnosis and co-morbidities	T3-12
	ASIA score (if known)	A (10)
	Time since injury (years)	> 1.5
Method of muscle strength test used	Manual Muscle Test	
	Isokinetic Dynamometry	Cybex Norm II speed 60 degrees/sec
	Hand Held Dynamometry	
Trained examiner (yrs of experience)		2 physios (1 with 16 years SCI experience) (1 with 6 years orthopaedic experience)
Blinded examiner?		blinded participants
Clinimetric domain evaluated	Reliability	inter- and intra-tester reliability
	Validity	
	Responsiveness	
	Interpretability	
Statistical analysis used		ICC (> 0.8 considered significant) p values <0.05 considered significant Cv SD of <10% considered reproducible SRD SEM
Results		movt ICC isom Conc Ecc (60deg/sec) elb flex .56 .48 .67 elb ext .62 .65 .74 sh add .71 .73 .60 sh abd .74 .75 .64 sh int .88 .88 .79 sh ext .88 .74 .86

			intra-	inter-	(CV SD%)
		movt	A1/A2	B1/B2	A1/B1
		elb flex isom	16.3	12.2	14.4
		elb flex conc	21.7	15.4	21.1
		elb flex ecc	12.1	12.4	15.7
		elb ext isom	14.2	8.4	13.1
		elb ext conc	12.2	10.6	11.4
		elb ext ecc	11.6	15.7	9.0
		sh add isom	4.7	6.5	7.2
		sh add conc	6.9	4.5	9.0
		sh add ecc	7.0	5.9	7.8
		sh abd isom	11.1	7.8	15.2
		sh add conc	8.6	11.1	19.2
		sh add ecc	16.0	8.6	19.4
		sh int isom	4.4	6.4	7.4
		sh int conc	6.6	5.3	9.6
		sh int ecc	7.2	6.2	10.6
		sh ext isom	8.7	8.1	6.7
		sh ext conc	12.7	13.4	15.5
		sh ext ecc	9.4	8.6	6.4
Auth ors Conclusions		In subjects with SCI isokinetic strength scores of elbow flexion and extension are not reproducible. Isometric contractions of the shoulder show most intra-tester reliability as the exact repositioning of the ID was guaranteed			
Reviewers Comments		small sample			
COSMIN score		poor: sample size good: assumable pts were stable			

		Study 5
Authors, year		Noreau, L, Vachon, J. 1998 Comparison of three methods to assess muscular strength in individuals with spinal cord injury
Muscles assessed in study		Elbow flexors/ extensors shoulder abductors/adductors shoulder flexors/extensors
Types of participants including diagnosis specifics	Age (years)	14-63
	Gender	31 men 7 women
	No. of participants	38
	Diagnosis and co-morbidities	C5-L3 23 paraplegics 15 tetraplegics
	ASIA score (if known)	A (21), B (9), C (4), D(4)
	Time since injury	paraplegic 1.6 +/- 0.7 months tetraplegics 2.1 +/- 2.1 months
Method of muscle strength test used	Manual Muscle Test	10 grade modified MRC scale
	Isokinetic Dynamometry	CYBEX II 60 degrees/sec
	Hand Held Dynamometry	Penny and Giles
Trained examiner (yrs of experience)		trained physio
Blinded examiner?		?
Clinimetric domain evaluated	Reliability	yes
	Validity	yes
	Responsiveness	yes
	Interpretability	
Statistical analysis used		Spearman correlation coefficient $p < 0.05$ Pearson correlation co-efficient $p < 0.05$
Results		Spearman correlation co-efficient between MMT and HHD

		<table><tr><td></td><td>para</td><td>tetra</td></tr><tr><td>muscles</td><td>adm</td><td>d/c</td></tr><tr><td>elb flex</td><td>0.48</td><td>0.26</td></tr><tr><td>elb ext</td><td>0.46</td><td>0.55</td></tr><tr><td>sh flex</td><td>0.63</td><td>0.60</td></tr><tr><td>sh ext</td><td>0.44</td><td>0.49</td></tr><tr><td>sh abd</td><td>0.64</td><td>0.57</td></tr><tr><td>sh add</td><td>0.67</td><td>0.34</td></tr></table> <table><tr><td></td><td>para</td><td>tetra</td></tr><tr><td>muscles</td><td>adm</td><td>d/c</td></tr><tr><td>elb flex</td><td>0.76</td><td>0.75</td></tr><tr><td>elb ext</td><td>0.70</td><td>0.82</td></tr><tr><td>sh flex</td><td>0.89</td><td>0.89</td></tr><tr><td>sh ext</td><td>0.85</td><td>0.83</td></tr><tr><td>sh abd</td><td>0.73</td><td>0.82</td></tr><tr><td>sh add</td><td>0.81</td><td>0.90</td></tr></table> <p>Pearson correlation co-efficients between HHD and ID</p> <p>range of myometry scores within each MMT grade led to significant overlaps between 2 adjacent MMT grades, variable correlations between MMT and myometry moderate to strong correlations between myometry and ID</p>		para	tetra	muscles	adm	d/c	elb flex	0.48	0.26	elb ext	0.46	0.55	sh flex	0.63	0.60	sh ext	0.44	0.49	sh abd	0.64	0.57	sh add	0.67	0.34		para	tetra	muscles	adm	d/c	elb flex	0.76	0.75	elb ext	0.70	0.82	sh flex	0.89	0.89	sh ext	0.85	0.83	sh abd	0.73	0.82	sh add	0.81	0.90
	para	tetra																																																
muscles	adm	d/c																																																
elb flex	0.48	0.26																																																
elb ext	0.46	0.55																																																
sh flex	0.63	0.60																																																
sh ext	0.44	0.49																																																
sh abd	0.64	0.57																																																
sh add	0.67	0.34																																																
	para	tetra																																																
muscles	adm	d/c																																																
elb flex	0.76	0.75																																																
elb ext	0.70	0.82																																																
sh flex	0.89	0.89																																																
sh ext	0.85	0.83																																																
sh abd	0.73	0.82																																																
sh add	0.81	0.90																																																
Authors Conclusions		MMT does not appear to be sufficiently sensitive to determine improvements in muscle strength especially >grade 4. Recommend HHD to objectively measure muscle strength																																																
Reviewers Comments		only study to compare all 3 methods																																																
COSMIN score		reliability: poor measurements not independent content validity: fair, methodological flaws criterion validity: fair, sample size, methodological flaws Responsiveness: fair, sample size																																																

		study 6																									
Authors, year		Savic, G, Bergstrom, EMK, Frankel, HL, Jamous, MA, Jones, PW. 2007 Inter-rater reliability of motor and sensory examinations performed according to American spinal injury association standards																									
Muscles assessed in study		ASIA standards, all upper and lower limb muscles																									
Types of participants including diagnosis specifics	Age (years)	40.3																									
	Gender	38 men 7 women																									
	No. of participants	45 (43 completed motor initial exam; 22 completed test-re-test motor exam)																									
	Diagnosis and co-morbidities	15 C/sp, 29 T/sp, 1 L/sp																									
	ASIA score (if known)	A (24), B (4), C (4), D (13)																									
	Time since injury	3 months to 43 years																									
Method of muscle strength test used	Manual Muscle Test	0-5 MRC scale																									
	Isokinetic Dynamometry																										
	Hand Held Dynamometry																										
Trained examiner (yrs of experience)		yes trained clinical scientist and senior research physio																									
Blinded examiner?		No																									
Clinimetric domain evaluated	Reliability	Yes (intra-tester)																									
	Validity																										
	Responsiveness																										
	Interpretability																										
Statistical analysis used		ICC Kappa (MMT)																									
Results		inter-tester <table><tr><td>muscle</td><td>side</td><td>%</td><td>Kappa</td><td>weighted Kappa</td></tr><tr><td>biceps</td><td>R</td><td>91</td><td>0.694</td><td>0.694</td></tr><tr><td>biceps</td><td>L</td><td>91</td><td>0.730</td><td>0.649</td></tr><tr><td>ECR</td><td>R</td><td>86</td><td>0.752</td><td>0.932</td></tr><tr><td>ECR</td><td>L</td><td>91</td><td>0.823</td><td>0.973</td></tr></table>	muscle	side	%	Kappa	weighted Kappa	biceps	R	91	0.694	0.694	biceps	L	91	0.730	0.649	ECR	R	86	0.752	0.932	ECR	L	91	0.823	0.973
muscle	side	%	Kappa	weighted Kappa																							
biceps	R	91	0.694	0.694																							
biceps	L	91	0.730	0.649																							
ECR	R	86	0.752	0.932																							
ECR	L	91	0.823	0.973																							

		<table><tr><td>triceps</td><td>R</td><td>86</td><td>0.765</td><td>0.969</td></tr><tr><td>triceps</td><td>L</td><td>86</td><td>0.781</td><td>0.972</td></tr><tr><td>FDP</td><td>R</td><td>82</td><td>0.728</td><td>0.975</td></tr><tr><td>FDP</td><td>L</td><td>73</td><td>0.629</td><td>0.946</td></tr><tr><td>ADM</td><td>R</td><td>71</td><td>0.604</td><td>0.965</td></tr><tr><td>ADM</td><td>L</td><td>77</td><td>0.752</td><td>0.963</td></tr><tr><td>iliopsoas</td><td>R</td><td>73</td><td>0.593</td><td>0.963</td></tr><tr><td>iliopsoas</td><td>L</td><td>91</td><td>0.857</td><td>0.987</td></tr><tr><td>Quads</td><td>R</td><td>95</td><td>0.936</td><td>0.993</td></tr><tr><td>Quads</td><td>L</td><td>91</td><td>0.882</td><td>0.987</td></tr><tr><td>tib ant</td><td>R</td><td>77</td><td>0.670</td><td>0.972</td></tr><tr><td>tib ant</td><td>L</td><td>82</td><td>0.742</td><td>0.948</td></tr><tr><td>EHL</td><td>R</td><td>77</td><td>0.692</td><td>0.954</td></tr><tr><td>EHL</td><td>L</td><td>77</td><td>0.688</td><td>0.961</td></tr><tr><td>gastroc</td><td>R</td><td>77</td><td>0.707</td><td>0.943</td></tr><tr><td>gastroc</td><td>L</td><td>73</td><td>0.664</td><td>0.951</td></tr></table> <table><tr><td colspan="7">overall assignment of MMT grade % agreement</td></tr><tr><td>side</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr><tr><td>R</td><td>97</td><td>80</td><td>50</td><td>22</td><td>77</td><td>89</td></tr><tr><td>L</td><td>99</td><td>62</td><td>55</td><td>53</td><td>83</td><td>93</td></tr></table> average MMT percentage agreement between examiners left side 84% right side 82% strongest agreement grades 0 and 5 weakest agreement grades 2 and 3	triceps	R	86	0.765	0.969	triceps	L	86	0.781	0.972	FDP	R	82	0.728	0.975	FDP	L	73	0.629	0.946	ADM	R	71	0.604	0.965	ADM	L	77	0.752	0.963	iliopsoas	R	73	0.593	0.963	iliopsoas	L	91	0.857	0.987	Quads	R	95	0.936	0.993	Quads	L	91	0.882	0.987	tib ant	R	77	0.670	0.972	tib ant	L	82	0.742	0.948	EHL	R	77	0.692	0.954	EHL	L	77	0.688	0.961	gastroc	R	77	0.707	0.943	gastroc	L	73	0.664	0.951	overall assignment of MMT grade % agreement							side	0	1	2	3	4	5	R	97	80	50	22	77	89	L	99	62	55	53	83	93
triceps	R	86	0.765	0.969																																																																																																										
triceps	L	86	0.781	0.972																																																																																																										
FDP	R	82	0.728	0.975																																																																																																										
FDP	L	73	0.629	0.946																																																																																																										
ADM	R	71	0.604	0.965																																																																																																										
ADM	L	77	0.752	0.963																																																																																																										
iliopsoas	R	73	0.593	0.963																																																																																																										
iliopsoas	L	91	0.857	0.987																																																																																																										
Quads	R	95	0.936	0.993																																																																																																										
Quads	L	91	0.882	0.987																																																																																																										
tib ant	R	77	0.670	0.972																																																																																																										
tib ant	L	82	0.742	0.948																																																																																																										
EHL	R	77	0.692	0.954																																																																																																										
EHL	L	77	0.688	0.961																																																																																																										
gastroc	R	77	0.707	0.943																																																																																																										
gastroc	L	73	0.664	0.951																																																																																																										
overall assignment of MMT grade % agreement																																																																																																														
side	0	1	2	3	4	5																																																																																																								
R	97	80	50	22	77	89																																																																																																								
L	99	62	55	53	83	93																																																																																																								
Authors Conclusions		very good levels of agreement between 2 experienced examiners in all components of ASIA neuro exam																																																																																																												
Reviewers Comments		study looks at reliability of the ASIA standards neuro exam, took the MMT component out for review																																																																																																												
COSMIN score		poor: sample size Good: assumable pt were stable																																																																																																												

		Study 7
Authors, year and ref.		Jonsson, M, Tollback, A, Gonzales, H, Borg, J. 2000 Inter-rater reliability of the 1992 international standards for neurological and functional classification of incomplete spinal cord injury
Muscles assessed in study		ASIA neuro exam all upper and lower limb muscles
Types of participants including diagnosis specifics	Age	?
	Gender	15 men 8 women
	No. of participants	23
	Diagnosis and co-morbidities	12 C/sp, 6 T/sp, 5 L/sp
	ASIA score (if known)	3 complete 20 incomplete
	Time since injury	> 1 week
Method of muscle strength test used	Manual Muscle Test	grade 0-5 MRC scale
	Isokinetic Dynamometry	
	Hand Held Dynamometry	
Trained examiner (yrs of experience)		2 physicians (1 with 1 yr, 1 with >10 yrs experience) 2 physio (both > 10 yrs)
Blinded examiner?		yes
Clinimetric domain evaluated	Reliability	yes inter-tester
	Validity	
	Responsiveness	
	Interpretability	
Statistical analysis used		Kappa
Results		Inter-tester reliability before and after standardisation of assessment level before after R L R L

		C5 0.3 high agree 0.57 0.75 C6 0.33 0.46 0.67 0.87 C7 0.87 0.46 0.47 0.46 C8 0.82 0.63 0.57 0.65 T1 0.76 0.86 0.62 0.47 L2 0.54 0.48 0.56 0.63 L3 0.7 0.55 0.7 0.72 L4 0.53 0.49 0.89 0.69 L5 0.51 0.53 0.66 0.52 S1 0.61 0.55 0.64 0.69
Authors Conclusions		weak inter-tester reliability for scoring incomplete SCI. Agreement improved for motor scores once assessment standardised
Reviewers Comments		Looking at standards rather than MMT specifically
COSMIN score		Poor: sample size Good

		study 8
Authors, year and ref.		Schwartz, S, Cohen, ME, Herbison GJ, Shah, A. 1992 Relationship between 2 measures of upper extremity strength
Muscles assessed in study		biceps, triceps and extensor carpi radialis
Types of participants including diagnosis specifics	Age	15-70
	Gender	122 men
	No. of participants	122
	Diagnosis and co-morbidities	tetraplegics C4,C5, C6
	ASIA score (if known)	A-D
	Time since injury	> 72 hours
Method of muscle strength test used	Manual Muscle Test	10 point modified MRC scale
	Isokinetic Dynamometry	
	Hand Held Dynamometry	Penny and Giles
Trained examiner (yrs of experience)		yes physician and technician
Blinded examiner?		?
Clinimetric domain evaluated	Reliability	reliability study of HHD conducted pre-study with an ICC of 0.85-0.95
	Validity	relationship between MMT and HHD
	Responsiveness	MMT v HHD for detecting strength improvements
	Interpretability	myometry translated to MMT
Statistical analysis used		Spearman rank correlation co-efficients linear regression analyses and scatter plots
Results		Spearman Rank Correlation between MMT and myometry time since injury muscle 72h 1w 1m 3m 6m 12m bicep L 0.86 0.84 0.68 0.82 0.59 0.42 bicep R 0.80 0.83 0.79 0.68 0.59 0.18 ECR L 0.92 0.86 0.81 0.84 0.84 0.77 ECR R 0.94 0.78 0.93 0.79 0.75 0.71 validity- 22/24 correlations were significant between

		HHD and MMT (0.59-0.94) responsiveness- myometry measurements detect increases over time not reflected by MMT interpretability - range in scores most specific <4 and less specific >4
Authors Conclusions		Range of HHD measurements for a particular MMT grade were better for grades <4 and less specific for grades < 4. Results indicate that HHD detect increases in strength over time which are not reflected by changes in MMT scores
Reviewers Comments		Good study and comments
COSMIN score		reliability: Good stats content validity: excellent criterion validity (usual practice): good responsiveness: excellent interpretability: fair

		Study 9																											
Authors, year, title		Aufsesser, P, Horvat, M, Austin, R. 2003 The reliability of hand held muscle testers with individuals with spinal cord injury																											
Muscles assessed in study		elbow flexors, extensors wrist extension																											
Types of participants including diagnosis specifics	Age	52 +/- 16 (25-83)																											
	Gender	25 men																											
	No. of participants	25																											
	Diagnosis and co-morbidities	C2-L4 14 tetraplegic 11 paraplegic																											
	ASIA score (if known)																												
	Time since injury	13 (1-34 years)																											
Method of muscle strength test used	Manual Muscle Test																												
	Isokinetic Dynamometry																												
	Hand Held Dynamometry	make technique MicroFet																											
Trained examiner (yrs of experience)		university professor experienced in HHD physio with no HHD experience																											
Blinded examiner?		?																											
Clinimetric domain evaluated	Reliability	inter and intra tester																											
	Validity																												
	Responsiveness																												
	Interpretability																												
Statistical analysis used		ICC root mean square																											
Results		<table><tr><td>Muscle</td><td></td><td>intra-</td><td>ICC</td><td></td><td>inter-</td></tr><tr><td></td><td>Av 1</td><td>single 1</td><td>Av 2</td><td>single 2</td><td>av</td><td>single</td></tr><tr><td>elb flex L</td><td>.93</td><td>.82</td><td>.98</td><td>.95</td><td>.36</td><td>.22</td></tr><tr><td>elb flex R</td><td>.99</td><td>.96</td><td>.96</td><td>.90</td><td>.21</td><td>.11</td></tr></table>	Muscle		intra-	ICC		inter-		Av 1	single 1	Av 2	single 2	av	single	elb flex L	.93	.82	.98	.95	.36	.22	elb flex R	.99	.96	.96	.90	.21	.11
Muscle		intra-	ICC		inter-																								
	Av 1	single 1	Av 2	single 2	av	single																							
elb flex L	.93	.82	.98	.95	.36	.22																							
elb flex R	.99	.96	.96	.90	.21	.11																							

		elb ext L .98 .94 .99 .97 .89 .80 elb ext R .96 .88 .99 .97 .74 .59 wrist ext L .98 .96 .99 .97 .84 .73 wrist ext R .98 .94 .99 .97 .84 .72
Authors Conclusions		intra-tester showed good reliability whereas inter-tester was not reliable or practical in a clinical setting. Authors recommend fixed lever arm instead of testers hand to increase reliability
Reviewers Comments		gaps in study design
COSMIN score		Poor: sample size Fair

		study 10																								
Authors, year, title		Herbison, GJ, Isaac, Z, Cohen, ME, Ditunno, JF 1996 Strength post-spinal cord injury: myometer vs manual muscle test																								
Muscles assessed in study		elbow flexors																								
Types of participants including diagnosis specifics	Age	34 (15-68)																								
	Gender	78 men 10 women																								
	No. of participants	88																								
	Diagnosis and co-morbidities	C4-C8																								
	ASIA score (if known)	A-D																								
	Time since injury	< 72 hours																								
Method of muscle strength test used	Manual Muscle Test	10 point modified MRC scale																								
	Isokinetic Dynamometry																									
	Hand Held Dynamometry	Penny and Giles																								
Trained examiner (yrs of experience)		yes 6 physicians, 3 technicians																								
Blinded examiner?		?																								
Clinimetric domain evaluated	Reliability	inter-tester reliability for MMT and myometer pre-testing (.94 and .82 respectively)																								
	Validity	myometer vs MMT																								
	Responsiveness	myometer vs MMT, change of scores																								
	Interpretability																									
Statistical analysis used		Students t-test co-efficients of variation																								
Results		<table> <thead> <tr> <th>MMT</th> <th>mean geometric %</th> <th>p value</th> </tr> </thead> <tbody> <tr> <td>3.5 =</td> <td>116</td> <td>>0.05</td> </tr> <tr> <td>3.5 to 4.0</td> <td>205</td> <td><0.002</td> </tr> <tr> <td>3.5 to 4.5</td> <td>232</td> <td><0.05</td> </tr> <tr> <td>4.0 =</td> <td>140</td> <td><0.05</td> </tr> <tr> <td>4.0 to 4.5</td> <td>139</td> <td><0.02</td> </tr> <tr> <td>4.0 to 5.0</td> <td>191</td> <td><0.001</td> </tr> <tr> <td>4.5 =</td> <td>127</td> <td>< 0.01</td> </tr> </tbody> </table>	MMT	mean geometric %	p value	3.5 =	116	>0.05	3.5 to 4.0	205	<0.002	3.5 to 4.5	232	<0.05	4.0 =	140	<0.05	4.0 to 4.5	139	<0.02	4.0 to 5.0	191	<0.001	4.5 =	127	< 0.01
MMT	mean geometric %	p value																								
3.5 =	116	>0.05																								
3.5 to 4.0	205	<0.002																								
3.5 to 4.5	232	<0.05																								
4.0 =	140	<0.05																								
4.0 to 4.5	139	<0.02																								
4.0 to 5.0	191	<0.001																								
4.5 =	127	< 0.01																								

		4.5 to 5.0 126 <0.02
Authors Conclusions		results of study suggest that increases in strength identified by myometric measurements were masked by MMT
Reviewers Comments		only looking at strength scores grade 3.5 or above on MMT pre-study
COSMIN score		content validity: fair (flaws) criterion validity (MMT usual practice): poor sample size, fair responsiveness: good

		Study 11
Authors, year and ref.		Burns, SP, Spanier, DE. 2005 Break-technique hand held dynamometry: relation between angular velocity and strength measurements
Muscles assessed in study		elbow flexors and extensors
Types of participants including diagnosis specifics	Age	57 +/- 15 years
	Gender	19 men 1 woman
	No. of participants	20
	Diagnosis and co-morbidities	tetraplegia
	ASIA score (if known)	A (6), B (4), C (4), D (6)
	Time since injury	19 +/- 21 years
Method of muscle strength test used	Manual Muscle Test	MRC scale
	Isokinetic Dynamometry	
	Hand Held Dynamometry	Make and Break technique Chatillon CSD-200 dynamometer with an electro-goniometer to measure speeds of movement 15 (low), 33 (medium) and 55 (high) degrees/second
Trained examiner (yrs of experience)		yes
Blinded examiner?		?
Clinimetric domain evaluated	Reliability	test - retest
	Validity	Yes comparison of strength scores using make and break technique.
	Responsiveness	
	Interpretability	
Statistical analysis used		ICC, Spearman correlation coefficient break/make ratio
Results		test-retest p value

		make 0.09 low break 0.51 medium break 0.04 high break 0.007 p value comparison sess 1 sess 2 make v low break 0.03 0.004 low break v med break 0.007 0.09 med break v high break 0.006 0.045 B/M ratios mean lowB/M 1.16 1.16 med B/M 1.3 1.33 high B/M 1.51 1.43 ICC for all testing techniques were >0.89
Authors Conclusions		Findings confirmed the authors hypothesis that greater angular velocity during break technique HHD testing is associated with higher strength measures. Variability in angular velocity is a potential source of error. To minimise variability in strength measurements maybe better to use make technique
Reviewers Comments		Looking at comparing make and break techniques at different speeds, only study to use movement during break technique (isotonic v isometric)
COSMIN score		Poor: sample size, stats Good

Abbreviations: musc, muscle; ext, shoulder external rotation; int, shoulder internal rotation; all, results for paraplegia and tetraplegia combined; para, paraplegia; tetra, tetraplegia; h, hours; w, week; m, months; adm, admission; d/c, discharge; elb, elbow; sh, shoulder; L, left; R, right; ex, extensors ; flex, flexors, add, adductors; abd, abductors; before and after relates to standardisation of procedure for testing muscle strength; ECR: extensor carpi radialis/ wrist extensors; ?, not stated, isom, isometric; sess, session.

Appendix V: Excluded Studies

1. Marino RJ, Jones L, Kirshblum S, Tal J, Dasgupta A. Reliability and repeatability of the motor and sensory examination of the international standards for neurological classification of spinal cord injury. J Spinal Cord Med.2008; 31: 166-170.

– *excluded as study looked at total motor scores rather than individual MMT scores*

2. Cohen ME, Ditunno JF, Donovan WH, Maynard FM. A test of the 1992 international standards for neurological and functional classification of spinal cord injury. Spinal cord.1998; 36(): 554-560.

– *excluded as study looked at total motor scores rather than individual MMT scores*