Meeta, Poonam and Claydon, Leica S. and Hendrick, Paul and Cook, Chad and Baxter, David G. (2016) Pain and physical functioning in neuropathic pain: a systematic review of psychometric properties of various outcome measures. Pain Practice, 16 (4). pp. 495-508. ISSN 1533-2500

PAIN Practice

Pain and physical functioning in neuropathic pain: a systematic review of psychometric properties of various outcome measures

Journal:	*Pain Practice*
Manuscript ID:	PPR-2014-0207.R1
Manuscript Type:	Review
Keywords:	Assessment,Pain, Nerve Pain, Polyneuropathy

ABSTRACT

Introduction: A range of outcome measures across various domains are used to evaluate change following an intervention in clinical trials on chronic Neuropathic pain (NeP). However, in order to capture a real change in the variable of interest, the psychometric properties of a particular measure should demonstrate appropriate methodological quality. Various outcome measures in the domains of pain and physical functioning have been used in the literature for NeP, for which individual properties (e.g., reliability/validity) have been reported. To date, there is no definitive synthesis of evidence on the psychometric properties of those outcome measures, thus the aim of this systematic review was to evaluate the methodological quality [COnsensus based Standards for the selection of health status Measurement INstruments (COSMIN) guidelines] of studies that evaluated psychometric properties of pain and physical functioning outcome measures used for NeP.

Methods: Specific MeSH/key-words related to three areas (pain and/or physical functioning, psychometric properties, and NeP) were used to retrieve relevant studies (English language) in key electronic databases (Medline (Ovid), CINAHL (EBSCO), Scopus, AMED and Web of Science) from database inception- July 2012. Articles retrieval/screening and quality analysis (COSMIN) were carried out by two independent reviewers.

Results: 24 pain and 37 physical functioning outcome measures were identified, varying in methodological quality from Poor-Excellent.

Conclusion: Although a variety of pain and physical functioning outcome measures have been reported in the literature, few have demonstrate methodologically strong psychometric properties. Thus, future research is required to further investigate the psychometric properties of existing pain and physical functioning outcome measures used for clinical and research purposes.

Keywords: neuropathic pain; systematic review; pain; physical function; outcome measures; psychometric properties; reliability; validity; responsiveness

1. INTRODUCTION

Neuropathic pain (NeP) is defined by the International Association for the Study of Pain's Neuropathic Pain Special Interest Group (NeuPSIG) as "*pain arising as a direct consequence of a lesion or disease affecting the somatosensory system*".[1] A range of assessment guidelines have been developed from the Initiative on Methods, Measurement and Pain Assessment in Clinical Trials (IMMPACT),[2] the European Federation of Neurological Societies (EFNS),[3] and the NeuPSIG[4] for NeP clinical trials and for clinical practice. These guidelines advocate a range of measures for assessing the core domains of pain, quality of life, mood, sleep, and functional capacity (physical, cognitive, emotional, and social). This notwithstanding, a variety of outcome measures are available for the above stated domains.[2] In order to evaluate the applicability of these measures, a systematic review of psychometric properties of available outcome measures used in published trials may provide a useful basis for selecting the best measurement instrument for a specific purpose.[5,6]

Individual assessment of psychometric properties of available outcome measures is important.[7,8] As part of this, in reviewing the evidence on available outcome measures, it is important to assess the methodological quality of those studies that investigated psychometric properties.[9] While in clinical practice adoption of outcome measures will depend on feasibility of use (speed, ease of use, and limited need for an overly sophisticated instrument),[10] emphases should be also be given to measures which are proven to be reliable, valid, and responsive/interpretable for a given population.

Pain remains a leading cause of disability at the individual level, associated with functional losses as well as mood disturbances.[11] Thus the focus of this systematic review will be in evaluating the psychometric properties of various outcome measures used in the domains of pain and physical functioning in NeP. On examination of the literature, a number of outcome measures have been identified in which have been used to measure pain intensity and physical function in NeP trials;[5,7,8,12] however, there is limited conclusive evidence on their psychometric properties. Use of reliable and valid outcome measures can help to better evaluate the patient's outcomes in terms of

pain and physical functioning, enabling better management, including the earliest appropriate management to minimize risks of co morbidities and disabilities.

Existing evidence on the psychometric properties of pain and physical functioning outcome measures used in NeP trials have not previously been systematically reviewed. The aim of this systematic review was to systematically review and identify the gaps in literature for the evaluated psychometric properties (reliability, validity, responsiveness, and interpretability) of identified outcome measures for 'pain and physical functioning' as recommended by the IMMPACT guidelines in NeP population. This review involved a systematic search of the literature. The findings of the current study may assist in outlining the effective intervention strategies for patients with NeP. The objectives of this systematic review were:

- Systematically review and identify the type of established psychometric properties for the identified outcome measures quantifying pain and physical functioning in neuropathic pain populations.
- Evaluate the methodological quality of the included studies investigating the psychometric properties of the identified outcome measures in the domain of pain and physical functioning in neuropathic pain populations in accordance with the Consensus-based Standards for the selection of Health Measurement Instruments (COSMIN) checklist with 4-point scale.

2. METHOD

2.1 Information sources

A systematic search was conducted following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. The following electronic databases were searched: Ovid Medline, CINAHL, Scopus, AMED, and Web of Science (WOS) (from database inception to 31st

July 2012). The search update engine from the available databases was activated in order to be familiar with the new searches in the current field, since the original search.

2.2 Search strategy

The key words and MESH headings in three broad areas (pain and/or physical functioning outcome measures, psychometric properties, and NeP) were used in the development of a search strategy (Table I). Several strategies were used to develop a comprehensive list of keywords/MeSH terms/subject headings representing each area. For outcome measures, all pain and physical functioning outcome measures that were used in clinical trials of NeP were chosen. For psychometric properties, we chose the standardised terminologies used by the COSMIN frame work.[6] For the terms relating to NeP, MESH terms/ key words indexed for neuropathy, neuralgia, and neurodynia were used. Words within each theme were combined with OR and across themes with AND. This search strategy was amended for different databases as necessary.

Insert Table I about here.

2.3 Study selection

Articles identified in the search underwent a series of screening processes. Firstly, duplicate articles were removed. Two reviewers (PM and LC) independently selected and screened articles for potential eligibility at the title and abstract stages. Full text articles of all potentially eligible abstracts were retrieved for application of the eligibility criteria. Disagreements between the reviewers regarding inclusion of individual studies were discussed during a consensus meeting and, when unresolved, were resolved by discussion with other reviewers (PH, CC, and GDB). References of the selected papers were further explored for relevant articles.

2.4 Eligibility criteria

Cross sectional studies and longitudinal cohort studies, which included at least one assessment of a psychometric property of a pain or functional outcome measure in a NeP population (Nep as defined by the Clinical Resource Efficiency Support Team- CREST)[13] were included. The adopted search strategy revealed two distinct categories of evaluations: one intended for screening or diagnosis, and the other developed to measure outcomes. Since the focus of this review was to investigate the psychometric properties of tools used to measure changes in the status of either pain or functional outcomes over time: screening or diagnostic tools were excluded. Studies published as case report, editorial, or reviews were also excluded. Only articles published in the English language and on humans were selected.

2.5 Data extraction and synthesis

A systematic approach to data extraction was carried out by independent reviewers (PM and LC/ PH/ CC/ GDB), with equal number of articles randomly distributed among the team members. Each member extracted the data from the allotted articles, which were then checked for accuracy, with consensus meetings and opinions from other reviewers to resolve any disagreements. The following data were collected and tabulated from each of the included articles: study reference, participant characteristics, outcome measures studied, and type of psychometric properties tested (reliability and/or validity) (Table II). Further summary of identified outcome measures with their published psychometric properties and COSMIN grading were synthesized (Table IV & V). Results from excellent and good methodological quality studies based on COSMIN criteria (as stated in Table VI) were used to formulate recommendations for acceptable psychometric properties scores (for definitions of acceptable, good and excellent scores see Table VI).

2.6 Methodological quality of individual studies reporting on psychometric properties

Whereas a variety of tools are available to measure the methodological quality of studies that report on scale development and assessed psychometric properties, the Consensus-based Standards

for the selection of Health Measurement Instruments (COSMIN)[6] checklist; developed by an international group of experts, is unique and preferred because it allows for individual assessment of each psychometric domain within a study.

The COSMIN checklist[14] (Table III) consists of 'A to J' nine boxes (Internal consistency-.Box A; Reliability- Box B; Measurement error- Box C; Content validity- Box D; Structural validity- Box E; Hypotheses testing- Box F; Cross-cultural validity- Box G; Criterion validity- Box H; Responsiveness- Box I; Interpretability- Box J), with 5–18 items concerning methodological standards for how each measurement property should be assessed. According to COSMIN guidelines, the methodological quality of a study is considered adequate if all items in a box (A to J) were considered adequate. For this, each item was scored on a 4-point rating scale (i.e., ''poor'', ''fair'', ''good'', or ''excellent''). The primary investigator (PM) independently scored all articles and the results were discussed and consensus obtained with each relevant team member. Methodological quality was determined using the 'lowest rating score'[6] achieved by any item for the representative psychometric property. Therefore, if one criterion for any property scored 'poor', the methodological quality for that particular property was rated as 'poor' overall, irrespective of the scores that other criteria achieved. Disagreements regarding COSMIN scoring were resolved by discussion between reviewers. Reviewers were not blinded to the journal affiliation or authors of the included articles.

Insert Table III about here.

3. RESULTS

Figure 3.1 illustrates the study selection process. The search resulted in 10,913 articles. After accounting for duplicate removal, title screening, and abstract screening, 80 articles were identified and retrieved as potentially eligible for the review. While checking the eligibility of full text articles, a further 16 articles were excluded from the review as two articles were editorial papers; two were commentary papers; five articles were based on cancer pain; three papers were PhD publications;

and for the remaining four, full text article were not available. Thus total of 64 articles satisfied our eligibility criteria and were included in this review.

Insert Figure I about here

3.2 Characteristics of included studies

In total, 64 studies reporting 61 different outcome measures were identified. The included studies evaluated the psychometric properties of pain outcome domains (n=24) and physical function outcome domains (n= 37), (Table II). For the 24 pain intensity outcome measures, fifteen (63%), measures were patient-reported/self-reported measures, and the rest nine (37%) were the therapist/ clinician completed measures. For the 37 physical function outcome measures, seventeen (46%) measures were patient-reported/ self-reported measures i.e. symptomatic assessment (subjective), nine (24%) measures were performance based measures, and the rest of the eleven (30%) measures were therapist completed measures i.e. symptoms and signs (subjective and objective testing). The synthesis of results per/ outcome measure, their published psychometric properties, and quality assessment scores for studies, are detailed in Table IV and V. Data on the characteristics of the study population and sample population were extracted on the interpretability and generalizability boxes provided by the COSMIN checklist. Information regarding the sample size and gender distribution is reported in Table II.

Insert Table II about here.

3.2.1 Pain intensity outcome measures

Pain domain outcomes (Table II, and IV) included: Brief Pain Inventory Scale for Diabetic Peripheral Neuropathy;[15] Complex Regional Pain Syndrome Severity Score;[16] Diabetes Symptom Checklist Type-2;[17] Foot Function Index (pain subscale);[18] Italian Neuropathic Pain Symptom Inventory;[19] McGill Pain Questionnaire;[20] modified Toronto Clinical Neuropathy Score;[21] Neuropathic Pain Scale;[22-24] Neuropathic Pain Sensory Inventory;[25,26] 0-10 Numerical Rating Scale;[27] Neuropathy

Total Symptom Score-6;[28] 0-10 point Pain Intensity- Numerical Rating Scale;[29] Pain Quality Assessment Scale;[30,31] Portuguese version of the Neuropathic Pain Symptoms Inventory;[32] Quantitative Sensory Testing (hot and cold pain threshold);[33-35] Sensory evaluation with Semmens-Weinstein Monofilaments;[36] Short-form McGill Pain Questionnaire-2;[37] Spanish Neuropathic Pain Symptom Inventory;[38] Toronto Clinical Scoring System;[39] Total Neuropathy Score;[40] Trauma Related Neuronal Dysfunction Symptoms Inventory;[41] Utah Early Neuropathy Scale;[42] Visual Analog Scale;[43] and Zoster Brief Pain Inventory.[44,45]

3.2.2 Physical functioning outcome measures

The range of physical functioning outcome measures was equally extensive, and included (Table II, and V): Alderson-McGall Hand Function questionnaire;[46] Barthel Index;[47] Berg Balance Measure;[48] Brief Pain Inventory Facial;[49] Charcot-Marie-Tooth disease Neuropathy score;[50,51] Charcot-Marie-Tooth disease Neuropathy Score-2;[52] Disabilities of Arm, Shoulder and Hand Questionnaire;[53-56] Deambulation Index;[47] Dellon-modified Moberg pick-up test;[57] Facial Disability Index;[58] Functional Dexterity test;[59] Human Activity Profile;[60] INCAT The Overall Disability Sum Score;[61] Inflammatory neuropathy Sensory Score;[62] Levine-Katz Questionnaire;[56] Michigan Hand Outcome Questionnaire;[53] modified Neuropathy Disability Score;[63] 10-Meter walking test;[48,64] Nine-Hole Peg test;[64] Neuropathy Impairment Score;[51] Overall Disability Sum Score;[65] Overall Neuropathy Limitations Scale;[64,66] Patient Evaluation Measure;[53] Physical Performance Measures (6 minute walk test, Timed up and go test);[67] Questionnaire Rising and Sitting down;[68] Radboud skills Questionnaire;[69] short form Screening of Activity Limitation and Safety Awareness Scale;[70,71] Step Activity Monitor;[72] Step Activity Monitor (4 min walk test);[73] Sheehan Disability Scale;[74] Sollerman Hand function test;[59] Turkish version of the Boston Questionnaire;[75] Ulnar Neuropathy at the Elbow Questionnaire;[76] 12-Item Multiple Sclerosis Walking Scale;[77] Walking Stairs Questionnaire;[68] Work stimulation tasks (knob turn, Linear motion, and Lever arm);[78] and Zoster Impact Questionnaire.[45]

3.3 Methodological quality of studies evaluating psychometric properties of pain intensity and physical functioning outcome measures

3.3.1 Reliability

The majority of the instruments included in our review were not tested for all psychometric properties listed on COSMIN checklist. Forty four of the sixty four studies (68%) assessed various forms of reliability (Internal consistency, inter-rater reliability, intra-rater reliability, test-retest reliability, and measurement error) and showed a mixed methodological quality of evidence (excellent/good/fair/poor), when evaluated on COSMIN (Table IV and V). The key results for reliability showed that the BPI-DPN, and the SF-MPQ2 have excellent (α> 0.90) internal consistency. The mTCNS has good internal consistency (α= 0.81- 0.90), inter-rater reliability, and intra-rater reliability (ICC or Κ= 0.81-0.90). The hot and cold pain thresholds on the QST have good inter-rater and test-retest reliability (ICC or Κ= 0.81-0.90). The Spanish NPSI has excellent internal consistency(α> 0.90) with good test-retest reliability(ICC or Κ= 0.81-0.90). Measurement error was the least reported form of reliability, and the TRNDSI had good test-retest reliability (ICC or Κ= 0.81-0.90) and measurement error (see Table IV). These measures with excellent and good psychometric properties scores also scored good/excellent on the COSMIN checklist (as according to COSMIN criteria stated in Table VI).

3.3.2 Validity

Validity was the more frequently tested psychometric property, in forty nine of sixty four studies (76%), there was face/content validity, structural validity, construct validity, criterion/concurrent validity, convergent validity, discriminative validity, hypothesis testing, and responsiveness. Similar to the findings for reliability, mixed methodological quality evidence (excellent/good/fair/poor) was found when evaluated on COSMIN (Table IV and V). The key results for validity showed that the NPSI, the SALSA, and the UNEQ have excellent content validity as there

were no concerns raised by the patients or experts regarding the wording of questionnaires, and thus no further modifications were advised. The UENS has the best criterion validity followed by the HAP and the mNDS. Approximately one third of the studies (18/49, 36%) evaluated responsiveness form of validity. The NPS has excellent responsiveness followed by the 0-10 PI NRS, and the ODSS. Also the studies showing these evidences were of excellent/good methodological quality on the COSMIN checklist (as according to COSMIN criteria stated in Table VI).

Insert Table IV and V about here.

4. Discussion

To our knowledge, this is the first systematic review to evaluate the evidence for the psychometric properties of pain and physical functional outcome measures used in assessment in NeP conditions, and to identify the methodological quality of the studies investigating the psychometric properties of various outcome measures. A total of 61 different outcome measures were identified related to the domains of pain and physical functioning. In this systematic review, while most of the studies have shown good/excellent evidence of reliability and validity of the used scales, only few are considered 'excellent to good' in terms of their methodological quality. Our review identified acceptable reliability and validity (for a few key properties) for the mTCNS, the TRNDI, the 0-10 PI NPS, the QST, the SALSA, the Spanish NPSI, the ODSS, the SF-MPQL, the UNEQ, the UENS, the HAP, the mNDS, the NDS and the BPI-DPN.

The available studies investigating the psychometric property of reliability were rated in varying methodological quality from 'poor' to 'excellent' on the COSMIN checklist. However, the majority of studies showed similar methodological shortcomings. In this review, smaller sample sizes were found to be associated with the majority of inconsistent results. According to COSMIN guidelines,[6] a sample size of ≥100 is considered to be an adequate/ excellent sample size, given the need for precision in the overall estimates; these estimates are based on the power 0.80.[79, 25] A

sample size of 50 provides a 0.70 power (level of significance being 0.05), while 100 has a power of 0.94.[25]

In the current systematic review, many outcome measures seem promising for different domains of reliability and validity (according to COSMIN criteria stated in Table VI), as the FFI, the NTSS-6, the AMHFQ, the DASH, the HAP, the ISS, the MHQ, the PEM, the SDS, the TBQ, the UNEQ, and the Walk-12 scales have 'moderate' (α> 0.71-0.80) to 'excellent' (α> 0.90) published grades for internal consistency. However, when the methodological quality of the studies were evaluated on COSMIN, these were graded of 'poor/fair' quality because of the small sample size. These findings are consistent with those of a recent systematic review on outcome measures in neck pain, where smaller sample sizes frequently led to poorer results.[80] This current review recommends that future research on a larger sample size (n= ≥100, as recommended by COSMIN) is needed to improve the quality of research on these measures.

Validity was the most frequently evaluated psychometric property in both pain and physical functioning outcome domains. The majority of these studies demonstrated unsatisfactory (poor/fair scores) results on COSMIN. The main reasons for this were inconsistencies in the following areas: smaller sample sizes; hypotheses were not formulated; and expected direction/magnitude of correlations was not stated in advance. Other common findings were a lack of information about reporting of missing items, and measures adopted to handle missing data. Though these two items did not contribute to the overall 'poor' grading on the COSMIN, it is expected that studies of 'good' methodological quality should report this construct, as a high number of missing items can introduce bias.

A further interesting finding of this review was that responsiveness was the least frequently studied psychometric property for the included pain and physical functioning outcome measures. There were a total of 18 studies which published the findings on responsiveness and only three scales- the NPS, the 0-10 PI NRS and the ODSS proved satisfactory methodological quality on

COSMIN. The remaining measures were graded 'fair to poor', and all the above stated shortcomings (small sample size, un-reporting of missing items, vagueness about how the missing data were handled, not well formulated hypothesis etc.) equally contributed to the inconsistent results for the studies reporting on this property.

In the current systematic review, there were few measures identified which had promising psychometric properties for key variables: the mTCNS (good internal consistency, inter-rater and intra-rater reliability and criterion validity); the TRNDSI, and the ZBPI (good test-retest reliability); the NPSI (excellent face/content validity); the 0 to 10 PI NRS (good responsiveness); the QST- pain threshold (good intra-rater and test-retest reliability); the NPS (excellent responsiveness); and the SALSA (excellent internal consistency and content validity), and were supported by a "excellent to good' methodological quality on the COSMIN checklist. The future use of these measures can be recommended based on their proven psychometric properties; however, it is imperative that other remaining psychometric properties of these outcome measures should also be established.

We also identified a list of instruments which showed their best methodological quality for few psychometric properties on COSMIN, but at the same time good methodological quality evidence was lacking for other properties: the TCSS (good construct validity, but poor inter and intra-rater reliability); the Short-form MPQ- 2 (excellent internal consistency, but fair construct validity and responsiveness); the HAP (good criterion validity, with poor internal consistency and responsiveness and fair hypothesis testing); the ODSS (good responsiveness but fair inter-rater and intra-rater reliability and construct validity); the UNEQ (excellent content validity, fair test-retest reliability, and poor internal consistency, construct validity, and responsiveness); the TBQ (good construct validity, fair test-retest reliability, and poor internal consistency); the UENS (excellent criterion validity, with poor inter-rater reliability and responsiveness); and the BPI-DPN (excellent internal consistency and discriminative validity, fair construct validity and poor criterion validity). Since study methodology may influence results for psychometric properties, it is recommended that

further evaluation of these psychometric properties with studies of improved methodological quality should be carried out.

Limitations

Firstly, it is acknowledged that 'Neuropathic Pain conditions' is an umbrella term which covers a range of different conditions such as diabetic neuropathy, trigeminal neuralgia, and post herpetic neuralgia.[81] For the search strategy, MESH terms/ key words indexed for neuropathy, neuralgia, and neurodynia were used to be as inclusive as possible. It is acknowledged that each condition could have been separately searched, and that such an approach may have lessened the chances of missing studies.

Secondly, psychometric properties such as reliability and validity, including responsiveness, are sub classified into various forms such as internal consistency, inter-rater/test retest reliability, content validity, minimal important difference, and standard error of measurement etc.[82] For the current search strategy, keywords in three broader areas (reliability and/or, validity and/or, and responsiveness) were used rather than individual sub classified keywords. However, since these broader terms are the most commonly used to denote the various forms of psychometric properties, it is anticipated that the majority of studies would have been selected.

Lastly, for this systematic review, multidisciplinary, international consensus-based methodological quality reporting guidelines, COSMIN, were followed for rating the quality of included studies of psychometric properties. The COSMIN checklist has well developed data extraction forms with detailed instructions for completion. The 4-point rating scale classifies each assessment of a measurement property as 'excellent, good, fair, or poor', based on the scores of the items in the corresponding COSMIN box. The methodological quality of a study is considered adequate if all items in a box (A to J) are considered adequate. However, frequently not all items in a box are scored adequate, and it is not feasible to provide overall definitive grade for each

psychometric property; thus no decisions can be drawn for the methodological quality of the studies based purely on COSMIN findings.

Conclusion

In this review we evaluated the evidence for psychometric properties of 61 unique outcome measures identified to assess pain and physical functioning outcome domains in trials of NeP conditions. We have presented extensive data which demonstrate the psychometric properties of these available outcome measures, and recommend the use of the mTCNS, the TRNDSI, the ZBPI, the NPSI, the 0 to 10 PI NRS, the QST- pain threshold, and the NPS to detect changes in pain intensity and physical functions. We found that important information regarding the methodological quality of the majority of studies demonstrating these psychometric properties is lacking or is of poor quality. Since NeP is a multi-disabling condition with significant associated morbidity, usage of quality evidenced pain and physical functional measures is a key recommendation for future research in NeP intervention studies. It appears that despite representing these measures in many studies of NeP, the methodological quality for most of the measures is not strong enough to recommend their use based on their psychometric properties. Thus, good quality future research is required to further investigate the psychometric properties of identified outcome measures used for clinical and research purposes.

ACKNOWLEDGEMENTS

All authors declare that there exist no conflicts of interest associated with the current study. The authors thank and acknowledge Dr Ramakrishnan Mani, Lecturer, Centre for Health Activity and Rehabilitation Research, School of Physiotherapy, University of Otago for his suggestions, invaluable constant assistance and helping with the constructive feedback on drafts of the manuscript. The findings of the study have been presented as a poster in the 8th Congress of the European Federation of IASP® Chapters (EFIC 2013) in Florence, Italy.

REFERENCES

1. Treede RD, Jensen TS, Campbell JN, et al. Neuropathic pain: redefinition and a grading system for clinical and research purposes. *Neurology.* Apr 29 2008;70(18):1630-1635.
2. Turk DC, Dworkin RH, Allen RR, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain.* Dec 2003;106(3):337-345.
3. Cruccu G, Sommer C, Anand P, et al. EFNS guidelines on neuropathic pain assessment: revised 2009. *European journal of neurology : the official journal of the European Federation of Neurological Societies.* Aug 2010;17(8):1010-1018.
4. Haanpaa M, Attal N, Backonja M, et al. NeuPSIG guidelines on neuropathic pain assessment. *Pain.* Jan 2011;152(1):14-27.
5. Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine.* Jan 2011;43(1):23-31.
6. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* Jul 2010;63(7):737-745.
7. Stinson JN, Kavanagh T, Yamada J, Gill N, Stevens B. Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents. *Pain.* Nov 2006;125(1-2):143-157.
8. Squires JE, Estabrooks CA, O'Rourke HM, Gustavsson P, Newburn-Cook CV, Wallin L. A systematic review of the psychometric properties of self-report research utilization measures used in healthcare. *Implement Sci.* 2011;6:83.
9. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a

scoring system for the COSMIN checklist. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.* May 2012;21(4):651-657.

10. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health technology assessment (Winchester, England).* 1998;2(14):i-iv, 1-74.

11. Gore M, Brandenburg NA, Hoffman DL, Tai KS, Stacey B. Burden of illness in painful diabetic peripheral neuropathy: the patients' perspectives. *The journal of pain : official journal of the American Pain Society.* Dec 2006;7(12):892-900.

12. Mehta P, Claydon L, Hendrick P, Winser S, Baxter GD. Outcome Measures in Randomized Controlled Trials of Neuropathic Pain Conditions: A Systematic Review of Systematic Reviews and Recommendations for Practice. *Clin J Pain.* Mar 21 2014.

13. Clinical Resource Efficiency Support Team (CREST). Guidelines on the management of neuropathic pain. 2008; http://www.thblack.com/links/RSD/CRESTManagementNeuropathicPainGuidelines.pdf. . Accessed Sept, 2012.

14. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology.* Jan 2007;60(1):34-42.

15. Zelman DC, Gore M, Dukes E, Tai KS, Brandenburg N. Validation of a modified version of the brief pain inventory for painful diabetic peripheral neuropathy. *J Pain Symptom Manage.* Apr 2005;29(4):401-410.

16. Harden RN, Bruehl S, Perez RS, et al. Development of a severity score for CRPS. *Pain.* Dec 2010;151(3):870-876.

17. Valk GD, Grootenhuis PA, van Eijk JT, Bouter LM, Bertelsmann FW. Methods for assessing diabetic polyneuropathy: validity and reproducibility of the measurement of sensory symptom severity and nerve function tests. *Diabetes Res Clin Pract.* Feb 2000;47(2):87-95.

18. Novak P, Burger H, Marincek C, Meh D. Influence of foot pain on walking ability of diabetic patients. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine.* Nov 2004;36(6):249-252.

19. Padua L, Briani C, Jann S, et al. Validation of the Italian version of the Neuropathic Pain Symptom Inventory in peripheral nervous system diseases. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology.* Apr 2009;30(2):99-106.

20. Helme RD, Katz B, Gibson S, Corran T. Can psychometric tools be used to analyse pain in a geriatric population? *Clin Exp Neurol.* 1989;26:113-117.

21. Bril V, Tomioka S, Buchanan RA, Perkins BA, m TSG. Reliability and validity of the modified Toronto Clinical Neuropathy Score in diabetic sensorimotor polyneuropathy. *Diabetic medicine : a journal of the British Diabetic Association.* Mar 2009;26(3):240-246.

22. Galer BS, Jensen MP. Development and preliminary validation of a pain measure specific to neuropathic pain: the Neuropathic Pain Scale. *Neurology.* Feb 1997;48(2):332-338.

23. Jensen MP, Dworkin RH, Gammaitoni AR, Olaleye DO, Oleka N, Galer BS. Assessment of pain quality in chronic neuropathic and nociceptive pain clinical trials with the Neuropathic Pain Scale. *The journal of pain : official journal of the American Pain Society.* Feb 2005;6(2):98-106.

24. Jensen MP, Friedman M, Bonzo D, Richards P. The validity of the neuropathic pain scale for assessing diabetic neuropathic pain in a clinical trial. *Clin J Pain.* Jan 2006;22(1):97-103.

25. Bouhassira D, Attal N, Fermanian J, et al. Development and validation of the Neuropathic Pain Symptom Inventory. *Pain.* Apr 2004;108(3):248-257.

26. Crawford B, Bouhassira D, Wong A, Dukes E. Conceptual adequacy of the neuropathic pain symptom inventory in six countries. *Health Qual Life Outcomes.* 2008;6:62.

27. Farrar JT, Pritchett YL, Robinson M, Prakash A, Chappell A. The clinical importance of changes in the 0 to 10 numeric rating scale for worst, least, and average pain intensity:

analyses of data from clinical trials of duloxetine in pain disorders. *The journal of pain : official journal of the American Pain Society.* Feb 2010;11(2):109-118.

28. Bastyr EJ, 3rd, Price KL, Bril V, Group MS. Development and validity testing of the neuropathy total symptom score-6: questionnaire for the study of sensory symptoms of diabetic peripheral neuropathy. *Clin Ther.* Aug 2005;27(8):1278-1294.

29. Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain.* Nov 2001;94(2):149-158.

30. Jensen MP, Gammaitoni AR, Olaleye DO, Oleka N, Nalamachu SR, Galer BS. The pain quality assessment scale: assessment of pain quality in carpal tunnel syndrome. *The journal of pain : official journal of the American Pain Society.* Nov 2006;7(11):823-832.

31. Jensen MP, Gould EM, Victor TW, Gammaitoni AR, White RE, Galer BS. The relationship of changes in pain quality to pain interference and sleep quality. *The journal of pain : official journal of the American Pain Society.* Aug 2010;11(8):782-788.

32. de Andrade DC, Ferreira KA, Nishimura CM, et al. Psychometric validation of the Portuguese version of the Neuropathic Pain Symptoms Inventory. *Health Qual Life Outcomes.* 2011;9(1):107.

33. Geber C, Klein T, Azad S, et al. Test-retest and interobserver reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): a multi-centre study. *Pain.* Mar 2011;152(3):548-556.

34. Maser RE, Nielsen VK, Bass EB, et al. Measuring diabetic neuropathy. Assessment and comparison of clinical examination and quantitative sensory testing. *Diabetes Care.* Apr 1989;12(4):270-275.

35. Felix ER, Widerstrom-Noga EG. Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. *J Rehabil Res Dev.* 2009;46(1):69-83.

36. Schreuders TA, Selles RW, van Ginneken BT, Janssen WG, Stam HJ. Sensory evaluation of the hands in patients with Charcot-Marie-Tooth disease using Semmes-Weinstein monofilaments. *Journal of hand therapy : official journal of the American Society of Hand Therapists.* Jan-Mar 2008;21(1):28-34; quiz 35.

37. Dworkin RH, Turk DC, Revicki DA, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain.* Jul 2009;144(1-2):35-42.

38. Villoria J, Rodriguez M, Berro MJ, Stern A, Sanchez-Magro I. Psychometric validation of the neuropathic pain symptom inventory for its use in Spanish. *J Pain Symptom Manage.* Jul 2011;42(1):134-146.

39. Bril V, Perkins BA. Validation of the Toronto Clinical Scoring System for diabetic polyneuropathy. *Diabetes Care.* Nov 2002;25(11):2048-2052.

40. Cornblath DR, Chaudhry V, Carter K, et al. Total neuropathy score: validation and reliability study. *Neurology.* Nov 10 1999;53(8):1660-1664.

41. Collins S, van Hilten JJ, Marinus J, Zuurmond WW, de Lange JJ, Perez RS. Development of a symptoms questionnaire for complex regional pain syndrome and potentially related illnesses: the Trauma Related Neuronal Dysfunction Symptoms Inventory. *Arch Phys Med Rehabil.* Jun 2008;89(6):1114-1120.

42. Singleton JR, Bixby B, Russell JW, et al. The Utah Early Neuropathy Scale: a sensitive clinical scale for early sensory predominant neuropathy. *Journal of the peripheral nervous system : JPNS.* Sep 2008;13(3):218-227.

43. Davidoff G, Morey K, Amann M, Stamps J. Pain measurement in reflex sympathetic dystrophy syndrome. *Pain.* Jan 1988;32(1):27-34.

44. Coplan PM, Schmader K, Nikas A, et al. Development of a measure of the burden of pain due to herpes zoster and postherpetic neuralgia for prevention trials: Adaptation of the brief pain inventory. *The Journal of Pain.* 2004;5(6):344-356.

45. Schmader KE, Sloane R, Pieper C, et al. The impact of acute herpes zoster pain and discomfort on functional status and quality of life in older adults. *Clin J Pain.* Jul-Aug 2007;23(6):490-496.

46. Alderson M, McGall D. The Alderson-McGall hand function questionnaire for patients with Carpal Tunnel syndrome: a pilot evaluation of a future outcome measure. *Journal of hand therapy : official journal of the American Society of Hand Therapists.* Oct-Dec 1999;12(4):313-322.

47. Padua L, Aprile I, Cavallaro T, et al. Relationship between clinical examination, quality of life, disability and depression in CMT patients: Italian multicenter study. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology.* Jun 2008;29(3):157-162.

48. Erdmann PG, van Meeteren NL, Kalmijn S, Wokke JH, Helders PJ, van den Berg LH. Functional health status of patients with chronic inflammatory neuropathies. *Journal of the peripheral nervous system : JPNS.* Jun 2005;10(2):181-189.

49. Lee JY, Chen HI, Urban C, et al. Development of and psychometric testing for the Brief Pain Inventory-Facial in patients with facial pain syndromes. *J Neurosurg.* Sep 2010;113(3):516-523.

50. Shy ME, Blake J, Krajewski K, et al. Reliability and validity of the CMT neuropathy score as a measure of disability. *Neurology.* Apr 12 2005;64(7):1209-1214.

51. Shy ME, Chen L, Swan ER, et al. Neuropathy progression in Charcot-Marie-Tooth disease type 1A. *Neurology.* Jan 29 2008;70(5):378-383.

52. Murphy SM, Herrmann DN, McDermott MP, et al. Reliability of the CMT neuropathy score (second version) in Charcot-Marie-Tooth disease. *Journal of the peripheral nervous system : JPNS.* Sep 2011;16(3):191-198.

53. Dias JJ, Rajan RA, Thompson JR. Which questionnaire is best? The reliability, validity and ease of use of the Patient Evaluation Measure, the Disabilities of the Arm, Shoulder and Hand and the Michigan Hand Outcome Measure. *J Hand Surg Eur Vol.* Feb 2008;33(1):9-17.

54. Eklund E, Svensson E, Hager-Ross C. Hand function and disability of the arm, shoulder and hand in Charcot-Marie-Tooth disease. *Disabil Rehabil.* 2009;31(23):1955-1962.

55. Novak CB, Anastakis DJ, Beaton DE, Mackinnon SE, Katz J. Relationships among pain disability, pain intensity, illness intrusiveness, and upper extremity disability in patients with traumatic peripheral nerve injury. *J Hand Surg Am.* Oct 2010;35(10):1633-1639.

56. Zimmerman NB, Kaye MB, Wilgis EF, Zimmerman RM, Dubin NH. Are standardized patient self-reporting instruments applicable to the evaluation of ulnar neuropathy at the elbow? *Journal of shoulder and elbow surgery / American Shoulder and Elbow Surgeons ... [et al.].* May-Jun 2009;18(3):463-468.

57. Amirjani N, Ashworth NL, Olson JL, Morhart M, Chan KM. Discriminative validity and test-retest reliability of the Dellon-modified Moberg pick-up test in carpal tunnel syndrome patients. *Journal of the peripheral nervous system : JPNS.* Mar 2011;16(1):51-58.

58. VanSwearingen JM, Brach JS. The Facial Disability Index: reliability and validity of a disability assessment instrument for disorders of the facial neuromuscular system. *Phys Ther.* Dec 1996;76(12):1288-1298; discussion 1298-1300.

59. Videler AJ, Beelen A, van Schaik IN, de Visser M, Nollet F. Manual dexterity in hereditary motor and sensory neuropathy type 1a: severity of limitations and feasibility and reliability of two assessment instruments. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine.* Feb 2008;40(2):132-136.

60. Farrell MJ, Gibson SJ, Helme RD. Measuring the activity of older people with chronic pain. *Clin J Pain.* Mar 1996;12(1):6-12.

61. Merkies IS, Schmitz PI. Getting closer to patients: the INCAT Overall Disability Sum Score relates better to patients' own clinical judgement in immune-mediated polyneuropathies. *Journal of neurology, neurosurgery, and psychiatry.* Aug 2006;77(8):970-972.

62. Merkies IS, Schmitz PI, van der Meche FG, van Doorn PA. Psychometric evaluation of a new sensory scale in immune-mediated polyneuropathies. Inflammatory Neuropathy Cause and Treatment (INCAT) Group. *Neurology.* Feb 22 2000;54(4):943-949.

63. Asad A, Hameed MA, Khan UA, Ahmed N, Butt MU. Reliability of the neurological scores for assessment of sensorimotor neuropathy in type 2 diabetics. *JPMA. The Journal of the Pakistan Medical Association.* Mar 2010;60(3):166-170.

64. Solari A, Laura M, Salsano E, Radice D, Pareyson D, Group C-TS. Reliability of clinical outcome measures in Charcot-Marie-Tooth disease. *Neuromuscul Disord.* Jan 2008;18(1):19-26.

65. Merkies ISJ. Clinimetric evaluation of a new overall disability scale in immune mediated polyneuropathies. *Journal of Neurology, Neurosurgery & Psychiatry.* 2002;72(5):596-601.

66. Graham RC, Hughes RA. A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. *Journal of neurology, neurosurgery, and psychiatry.* Aug 2006;77(8):973-976.

67. Manor B, Doherty A, Li L. The reliability of physical performance measures in peripheral neuropathy. *Gait Posture.* Aug 2008;28(2):343-346.

68. Perez R, Roorda LD, Zuurmond WWA, Bannink I, Vranken JH, de Lange JJ. Measuring perceived activity limitations in lower extremity Complex Regional Pain Syndrome type 1 (CRPS I): test–retest reliability of two questionnaires. *Clinical Rehabilitation.* 2002;16(4):454-460.

69. Oerlemans HM, Cup EH, DeBoo T, Goris RJ, Oostendorp RA. The Radboud skills questionnaire: construction and reliability in patients with reflex sympathetic dystrophy of one upper extremity. *Disabil Rehabil.* Mar 20 2000;22(5):233-245.

70. Melchior H, Velema J. A comparison of the Screening Activity Limitation and Safety Awareness (SALSA) scale to objective hand function assessments. *Disabil Rehabil.* 2011;33(21-22):2044-2052.

71. Group SCS, Ebenso J, Fuzikawa P, et al. The development of a short questionnaire for screening of activity limitation and safety awareness (SALSA) in clients affected by leprosy or diabetes. *Disabil Rehabil.* May 15 2007;29(9):689-700.

72. Smith DG, Domholdt E, Coleman KL, Del Aguila MA, Boone D. Ambulatory activity in men with diabetes: Relationship between self-reported and real-world performance-based measures. *Journal of rehabilitation research and development.* 2004;41(4):571.

73. van Schie CH, Noordhof EL, Busch-Westbroek TE, Beelen A, Nollet F. Assessment of physical activity in people with diabetes and peripheral neuropathy. *Diabetes Res Clin Pract.* Apr 2011;92(1):e9-11.

74. Rejas J, Pardo A, Ruiz MA. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *J Clin Epidemiol.* Apr 2008;61(4):350-356.

75. Sezgi¨n M, İncel NA, Sevi¨m S, Çamdevi¨ren H, As İ, ErdoĞan C. Assessment of symptom severity and functional status in patients with carpal tunnel syndrome: Reliability and validity of the Turkish version of the Boston questionnaire. *Disability & Rehabilitation.* 2006;28(20):1281-1286.

76. Mondelli M, Padua L, Giannini F, Bibbo G, Aprile I, Rossi S. A self-administered questionnaire of ulnar neuropathy at the elbow. *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology.* Dec 2006;27(6):402-411.

77. Graham RC, Hughes RA. Clinimetric properties of a walking scale in peripheral neuropathy. *Journal of neurology, neurosurgery, and psychiatry.* Aug 2006;77(8):977-979.

78. Kilmer DD, Aitkens SG, Wright NC, McCrory MA. Simulated work performance tasks in persons with neuropathic and myopathic weakness. *Arch Phys Med Rehabil.* Jul 2000;81(7):938-943.

79. Stevens J. *Applied multivariate statistics for the social sciences.* Taylor & Francis US; 2009.

80. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.* May 2012;21(4):659-670.

81. Dworkin RH, Backonja M, Rowbotham MC, et al. Advances in neuropathic pain: diagnosis, mechanisms, and treatment recommendations. *Arch Neurol.* Nov 2003;60(11):1524-1534.

82. Portney LG, Watkins MP. *Foundations of Clinical Research. Applications to Practice.* 3rd revised United States ed ed. Upper Saddle River/US: Prentice Hall; 2007.

FIGURE LEGENDS

Figure I Flow diagram summarising study selection process

Table I Search Strategy

Theme 1	**AND**	**Theme 2**	**AND**	**Theme 3**
Psychometric properties		**Pain and/or Physical functional Outcome measures**		**Neuropathic Pain**
Clinometric properties OR Validity OR Reliability OR Sensitivity OR Responsiveness OR Minimal(ly) clinically important difference OR Minimal(ly) clinically important change OR Minimum detectable change OR Smallest detectable change		Visual Analog Scale OR Numerical Pain Rating scale OR McGill pain rating scale OR Pain disability index OR Functional component of The Western Ontario and McMaster Universities Arthritis Index OR Timed scored functional activity OR Functional reach test OR Timed 9.1 metre to walk OR Disability of the arm shoulder and hand questionnaire OR Ulnar Neuropathy at Elbow questionnaire OR Daily activities by Verbal Rating Scale OR Function interference by Numerical Rating Scale		Pain OR Nerve pain OR Neuralgia OR Neurodynia OR Neuropathy

Table II Summary of included studies

Reference	Participant's characteristics		Outcome measures studied	Psychometric properties tested
Alderson & McGall 1999	Carpal Tunnel Syndrome	n= 17 Gender = 5 M, 12 F	Alderson-McGall hand function questionnaire	Reliability- Internal consistency, test-retest reliability; Validity- Convergent validity
Amirjani et al. 2011	Carpal Tunnel Syndrome	n= 162 Gender = 120 M, 42 F	Dellon-modified Moberg pick-up test	Reliability- test-retest reliability; Validity- Discriminative validity
Asad et al. 2010	Type 2 diabetics sensorimotor NeP	n= 60 Gender = not mentioned	modified Neuropathy Disability Score	Validity- Criterion validity
Bastyr et al. 2005	Diabetic peripheral NeP	n= 205 Gender = 122 M, 83 F	Neuropathy Total Symptom Score-6	Reliability- Internal consistency, test-retest reliability; Validity- Construct & Convergent validity, Responsiveness
Bouhassira et al. 2004	Peripheral and Central NeP	n= 176 Gender = 97 M, 79 F	Neuropathic Pain Symptom Inventory	Reliability- test-retest reliability; Validity- Face validity, Structural validity, Criterion validity, Convergent validity, Divergent validity, Responsiveness

Bril & Perkins 2002	Type 1 and 2 diabetic NeP	n= 89 Gender = 65 M, 24 F	Toronto Clinical Scoring System	Reliability- inter-rater, intra-rater reliability; Validity- Construct validity
Bril et al. 2009	Diabetic sensorimotor poly NeP	n= 65 Gender = 40 M, 25 F	modified Toronto Clinical Neuropathy Score	Reliability- Internal consistency, inter-rater, intra-rater reliability; Validity- Criterion validity
Collins et al. 2008	Complex regional pain syndrome-I	n= 27 Gender = 5 M, 22 F	Trauma Related Neuronal Dysfunction Symptoms Inventory	Reliability- test-retest reliability & Measurement error
Coplan et al. 2004	Herpes Zoster	n= 121 Gender = 45 M, 76 F	Zoster Brief Pain Inventory Questionnaire	Reliability- test-retest reliability; Validity- Hypothesis testing
Cornblath et al. 1999	Diabetic poly NeP	n= 30 Gender = 18 M, 12 F	Total Neuropathy Score	Reliability- inter-rater & intra-rater reliability; Validity- Construct validity
Crawford et al. 2008	Neuropathic Pain	n= 130 Gender = 70 M, 60 F	Neuropathic Pain Symptom Inventory questionnaire	Validity- Content validity
Davidoff et al. 1988	Reflex Sympathetic Dystrophy Syndrome	n= 17 Gender = 5 M, 12 F	Visual Analog Scales	Validity- Hypothesis testing
de Andrade et al. 2011	Neuropathic Pain	n= 94 Gender = 57 M, 37 F	Portuguese Neuropathic Pain Symptoms Inventory	Reliability- test-retest reliability; Validity- Face validity & Construct validity, Responsiveness

Dias et al. 2008	Wrist and hand disorders due to nerve involvement	n= 26 Gender = not mentioned	The Patient Evaluation Measure; The Michigan Hand Outcome Questionnaire; The Disabilities of Arm, Shoulder and Hand Questionnaire	Reliability- Internal consistency, test-retest Reliability; Validity- Construct validity
Dworkin et al. 2009	Diverse chronic pain syndrome; Diabetic NeP	n= 1108 Gender = 599 M, 509 F	Short-form McGill Pain Questionnaire- 2	Reliability- Internal consistency; Validity- Construct validity, Responsiveness
Eklund et al. 2009	Charcot-Marie-Tooth disease	n= 20 Gender = 9 M, 11 F	The Disabilities of Arm, Shoulder and Hand Questionnaire	Validity- Hypothesis testing
Erdmann et al. 2005	Chronic idiopathic demyelinating polyneuropathy; Multifocal Mono neuropathy	n= 30 Gender = 17 M, 13 F	Berg Balance Measure; 10 meter walk test	Validity- Hypothesis testing
Farrar et al. 2010	Diabetic peripheral NeP; Fibromyalgia syndrome	n= 1700 Gender = 680 M, 1020 F	0 to 10 Numeric Rating Scale	Validity- Responsiveness
Farrar et al. 2001	Diabetic peripheral NeP;	n= 984	0 to 10 point Pain Intensity	Validity- Responsiveness

	Post Herpetic Neuralgia	Gender = 567 M, 417 F	Numerical Rating Scale	
Farrell et al. 1996	Post Herpetic Neuralgia	n= 31 Gender = not mentioned	Human Activity Profile	Reliability- Internal consistency; Validity- Criterion validity & Hypothesis testing, Responsiveness
Felix & Widerstrom-Noga 2009	NeP related to Spinal Cord Injury	n= 22 Gender = 19 M, 3 F	Quantitative Sensory Testing (cold and heat pain thresholds)	Reliability- inter-rater & test-retest reliability; Validity- Construct validity
Galer & Jensen 1997	Post Herpetic Neuralgia; Diabetic NeP; Peripheral Nerve Injury	n= 160 (69; 24; 67) Gender = not mentioned	The Neuropathic Pain Scale	Validity- Hypothesis testing- Discriminative validity & Predictive validity
Geber et al. 2011	Peripheral Nerve lesion; Other neuropathies	n= 60 Gender = 37 M, 23 F	Quantitative Sensory Testing (heat, cold, mechanical and pressure pain threshold)	Reliability- inter-rater & test-retest reliability
Graham & Hughes 2006	Peripheral NeP	n= 65 Gender = 36 M, 29 F	12-Item Multiple Sclerosis Walking Scale	Reliability- Internal consistency & test-retest reliability; Validity- Hypothesis testing
Graham & Hughes 2006	Peripheral NeP	n= 100 Gender = 51:49	The Overall Neuropathy Limitations Scale	Reliability- Internal consistency, inter-rater, test-retest reliability;

				Validity- Content validity & Construct validity, Responsiveness
Harden et al. 2010	Complex and non-complex regional pain syndrome	n= 155 Gender = 68 M, 87 F	Complex regional pain syndrome severity score	Validity- Concurrent validity
Helme et al. 1989	Chronic Neuropathic Pain due to Post Herpetic Neuralgia	n= 49 Gender = 10 M, 39 F	McGill Pain Questionnaire	Validity- Concurrent validity
Jensen et al. 2005	Peripheral NeP	n= 133 Gender = 63 M, 70 F	The Neuropathic Pain Scale	Validity- Responsiveness
Jensen et al. 2006	Diabetes related foot pain	n= 159 Gender = 83 M, 76 F	The Neuropathic Pain Scale	Validity- Responsiveness
Jensen et al. 2006	Carpal Tunnel Syndrome	n= 40 Gender = 12 M, 2 F	Pain Quality Assessment Scale	Validity- Responsiveness
Jensen et al. 2010	Carpal Tunnel Syndrome	n= 100 Gender = 75 M, 25 F	Pain Quality Assessment Scale	Reliability- Internal consistency; Validity- Construct validity
Kilmer et al. 2000	Hereditary motor and sensory NeP	n= 9 Gender = 3 M, 6 F	Work stimulation tasks; Hand-held dynamometry	Reliability- test-retest reliability; Validity- Construct validity

Lee et al. 2010	Typical & atypical facial pain due to Trigeminal Neuralgia	n= 156 Gender = 58 M, 98 F	Brief Pain Inventory- Facial	Reliability- Internal consistency; Validity- Construct validity
Manor et al. 2008	Peripheral NeP	n= 20 Gender = 8 M, 12 F	Physical Performance Measures	Reliability- test-retest reliability
Maser et al. 1989	Diabetic neuropathy	n= 100 Gender = 54 M, 46 F	Quantitative sensory testing (thermal sensitivity)	Reliability- inter-rater reliability
Melchior & Velema 2011	Leprosy related Neuropathic Pain	n= 25 Gender = not mentioned	Screening of Activity Limitation and Safety Awareness Scale	Validity- Construct validity
Merkies & Schmitz 2006	Guillain Barré Syndrome; Chronic idiopathic demyelinating polyneuropathy	n= 20 Gender = 12 M, 8 F	The INCAT Overall Disability Sum Score	Validity- Concurrent validity
Merkies et al. 2002	Neuropathic Pain	n= 113 Gender = not mentioned	The Overall Disability Sum Score	Reliability- inter-rater & intra-rater reliability; Validity- construct validity, Responsiveness

Merkies et al. 2000	Neuropathic Pain	n= 113 Gender = not mentioned	Inflammatory Sensory Score	Reliability- Internal consistency, inter-rater, intra-rater reliability; Validity- Construct validity, Responsiveness
Mondelli et al. 2006	Ulnar Neuropathy at Elbow; Carpal Tunnel Syndrome	n= 292 Gender = 103 M, 189 F	Ulnar neuropathy at the elbow Questionnaire	Reliability- Internal consistency & test-retest reliability; Validity- content validity & construct validity, Responsiveness
Murphy et al. 2011	Charcot-Marie-Tooth disease	n= 34 Gender = not mentioned	Charcot-Marie-Tooth disease neuropathy score- 2	Reliability- inter-rater & intra-rater reliability
Novak et al. 2010	Peripheral Nerve injury	n= 124 Gender = 83 M, 41 F	The Disabilities of Arm, Shoulder and Hand Questionnaire	Reliability- Internal consistency; Validity- Construct validity
Novak et al. 2004	Type 2 diabetic NeP	n= 30 Gender = 10 M, 20 F	Foot Function Index (pain sub scale)	Reliability- Internal consistency; Validity- Hypothesis testing
Oerlemans et al. 2000	Reflex Sympathetic Dystrophy Syndrome	n= 54 Gender = 10 M, 44 F	The Radboud skills Questionnaire	Reliability- inter-rater & test-retest reliability; Validity- Construct validity
Padua et al. 2008	Charcot-Marie-Tooth disease	n= 211 Gender = 84 M, 127 F	Barthel Index; Deambulation Index	Validity- Construct validity

Padua et al. 2009	Peripheral Nerve disease	n= 392 Gender = 218 M, 174	Italian Neuropathic Pain Symptom Inventory	Reliability- test-retest reliability; Validity- Construct validity, Responsiveness
Perez et al. 2002	Complex regional pain syndrome-1	n= 21 Gender = 4 M, 17 F	Walking stairs Questionnaire; Questionnaire rising and sitting down	Reliability- test-retest reliability
Rejas et al. 2008	Neuropathic Pain	n= 603 Gender = 211 M, 392 F	Sheehan Disability Scale	Reliability- Internal consistency; Validity- Responsiveness
Schmader et al. 2007	Herpes Zoster	n= 165 Gender = 66 M, 99 F	Zoster Impact Questionnaire; Zoster Brief Pain Inventory	Validity- Hypothesis testing
Schreuders et al. 2008	Charcot-Marie-Tooth disease	n= 45 Gender = 25 M, 20 F	Sensory evaluation with Semmes-Weinstein Monofilaments	Validity- Construct validity
Sezgin et al. 2006	Idiopathic Carpal Tunnel Syndrome	n= 67 Gender = 5 M, 62 F	Turkish version of the Boston Questionnaire	Reliability- Internal consistency & test-retest reliability; Validity- Construct validity
Shy et al. 2005	Charcot-Marie-Tooth disease	n= 60 Gender = not mentioned	Charcot-Marie-Tooth disease neuropathy score	Reliability- Inter-rater & intra-rater reliability; Validity- Construct validity
Shy et al. 2008	Charcot-Marie-Tooth	n= 72	Charcot-Marie-Tooth disease	Validity- Responsiveness

	disease	Gender = 48 M, 24 F	Neuropathy Score; Neuropathy Impairment Score	
Singleton et al. 2008	Diabetic peripheral NeP	n= 129 Gender = not mentioned	The Utah Early Neuropathy Scale	Reliability- inter-rater reliability; Validity- Criterion validity, Responsiveness
Smith et al. 2004	Diabetic peripheral NeP	n= 57 Gender = 57 M, 0 F	Step Activity Monitor	Validity- Hypothesis testing
Solari et al. 2008	Charcot-Marie-Tooth disease	n= 40 Gender = 21 M, 19 F	The Overall Neuropathy Limitations Scale; 10 m walk; 9 hole peg test	Reliability- inter-rater & intra-rater reliability
The SALSA Group 2007	Leprosy &Diabetes related NeP	n= 568 Gender = 37.6%; 47% F	Screening of Activity Limitation and Safety Awareness Scale	Reliability- Internal consistency; Validity- Content validity
Valk et al. 2000	Type I and II Diabetes NeP	n= 78 Gender = 43 M, 35 F	The Diabetes symptom checklist-Type 2	Reliability- test-retest reliability; Validity- Construct validity
van Schie et al. 2011	Diabetic peripheral neuropathy	n= 24 Gender = 17 M, 7 F	Step Activity Monitor (4 minute walking test)	Validity- Construct validity & Criterion validity
VanSwearingen &	Facial paralysis	n= 46	Facial Disability Index	Reliability- Internal Consistency;

Brach 1996		Gender = 16 M, 30 F		Validity- Construct Validity
Videler et al. 2008	Hereditary motor and sensory type 1a neuropathy	n= 49 Gender = 21 M, 28 F	Sollerman Hand function test; Functional dexterity test	Reliability- Internal Consistency & test-retest reliability
Villoria et al. 2011	Chronic Neuropathic Pain	n= 548 Gender = 209 M, 339 F	Spanish Neuropathic Pain Symptom Inventory	Reliability- Internal Consistency, test-retest reliability; Validity- Construct validity
Zelman et al. 2005	Diabetic Peripheral NeP	n= 255 Gender = 114 M, 131 F	Brief Pain Inventory- Diabetic Peripheral Neuropathy scale	Reliability- Internal Consistency; Validity- Construct validity, Discriminative & Criterion validity
Zimmerman et al. 2009	Ulnar nerve injury	n= 48 Gender = not mentioned	The Disabilities of the Arm Shoulder and Hand Questionnaire; Levine-Katz Questionnaire	Validity- Criterion validity & Construct validity

Table III The COSMIN checklist with 4-point scale [Terwee 2012]

Step 1	Evaluated measurement properties in the article: Internal consistency, Reliability; relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability), Measurement error; absolute measures Content validity (including face validity), Structural validity, Hypothesis testing, Cross-cultural validity, Criterion validity, Responsiveness and Interpretability
Step 2	Determining if the statistical method used in the article are based on Classical Test Theory (CTT) or Item Response Theory (IRT): Box General requirements for studies that applied IRT models: excellent/ good/ fair/ poor
Step 3	Determining if a study meets the standards for good methodological quality: excellent/ good/ fair/ poor
Step 4	Determining the Generalizability of the results

Table IV Summary of identified pain Intensity outcome measures with their published psychometric properties and COSMIN grading

OMs	Reliability	**COSMIN**	Validity:	**COSMIN**	Responsiveness	**COSMIN**
BPI-DPN	***Internal consistency:*** *Zelman (2005):* BPI-DPN showed satisfactory unidimensionality both for the severity and the interference scales (Excellent, α= 0.94)	**excellent**	***Construct validity:*** *Zelman (2005):* BPI-DPN showed satisfactory construct validity for both the severity and the interference scales	**fair**	xx	xx
			Discriminant validity: *Zelman (2005):* Subcomponents of BPI-DPN: the severity and the interference scale showed satisfactory discriminant validity as both are correlated to a different extent with other measures- SF-12, and HADS (p< 0.001)	**excellent**		
			Criterion validity: *Zelman (2005):* BPI-DPN severity scale showed high and significant correlations with SF-12v2, and VRS, r's> 0.66 at p< 0.001	**poor**		

CRPS score	xx	xx	***Concurrent validity:*** *Harden (2010):* Higher CRPS scores were significantly associated with higher Rand 36 scores (pain intensity, worse physical and social functioning, greater role limitations due to physical and emotional problems, and lower energy and emotional well-being)	**fair**	xx	xx
dSCT2	***test-retest reliability:*** *Valk (2000):* Satisfactory test-retest correlation coefficient: severity of sensory alteration (0.89), and neuropathic pain (0.85)	**fair**	***Construct validity:*** *Valk (2000):* dSCT2 showed appropriate correlation with almost all nerve function tests	**fair**	xx	xx
FFI	***Internal consistency:*** *Novak (2004):* FFI pain subscale showed high unidimensionality (Excellent α= 0.9752)	**poor**	***Hypothesis testing:*** *Novak (2004):* FFI pain subscale showed moderate correlation with 6 meter walk test (r= -0.449, p< 0.001)	**fair**	xx	xx
Italian NPSI	***test-retest reliability:*** *Padua (2009):* Results showed high	**poor**	***Construct validity:*** *Padua (2009):* I-NPSI scores showed significant	**fair**	***Responsiveness:*** *Padua (2009):* I-NPSI scores	**fair**

	agreement between I-NPSI scores at two different visits		correlation with DN4, VAS and ID pain changes (p= 0.001)		represent reliable measurements to assess NeP symptoms and effectiveness of treatment on them	
MPQ	xx	xx	***Concurrent validity:*** *Helme (1989):* MPQ showed a significant correlation with VAS (r= 0.67), Word descriptor scale (r= 0.67), and ADL measures (r= 0.53, p< 0.001)	**poor**	xx	xx
mTCNS	***Internal consistency:*** *Bril (2009):* mTCNS showed satisfactory unidimensionality (Moderate, α= 0.78) ***inter-rater reliability:*** *Bril (2009):* Satisfactory ICC scores with good reliability (ICC= 0.83, 95% CI) ***intra-rater reliability:*** *Bril (2009):* Satisfactory correlation with	**good** **good** **good**	***Criterion validity:*** *Bril (2009):* Low but acceptable correlation with TCNS (Poor, γ= 0.58)	e**xcellent**	xx	xx

	symptom and sensory test (κ= 0.55-0.73)					
NPS	xx	xx	***Hypothesis testing: Descriptive validity-*** *Galer (1997):* 10 NPS pain descriptors showed minimal overlap between most items (γ< 0.50)	**poor**	***Responsiveness:*** *Jensen (2005):* NPS was significantly able to detect changes from pre-treatment to post treatment scores	**excellent**
			Predictive validity: *Galer (1997):* From 10 NPS pain descriptors, only four of descriptors (sharp, cold, sensitive and itchy pain) were able to discriminate PHN pain from other sources of pain, α= 0.01 level	**poor**	*Jensen (2006):* From 10 NPS pain descriptors, seven descriptors (intense, sharp, hot, dull, sensitive, unpleasant, and deep pain) were significantly able to pick up changes in score after treatment	**poor**
NPSI	***test-retest reliability:***	**fair**	***Face validity:***	**fair**	***Responsiveness:***	**poor**

Bouhassira (2004): Satisfactory ICC scores with excellent test retest reliability (ICC> 0.90)	*Bouhassira (2004):* The NPSI was completed accurately and appeared to be fully understood, notably by elderly subjects		*Bouhassira (2004):* Poor but acceptable correlations with PGIC and CGIC scores (ρ= 0:67; and ρ= 0.58)
	Content validity:	**excellent**	
	Crawford (2008): Majority of subjects did not raise any concerns with NPSI. Thus no changes to NPSI were consistently suggested		
	Structural validity:	**fair**	
	Bouhassira (2004): Each of five factors of NPSI corresponded to a relevant clinical component of NeP		
	Convergent validity:	**fair**	
	Bouhassira (2004): Poor but low correlation with global pain intensity measured by a numerical scale (ρ= 0.60, p< 0.001)		
	Divergent validity:	**fair**	
	Bouhassira (2004): No correlation with anxiety and depression scores measured by HADS (ρ= 0.27; and ρ=		

			0.32)			
			Criterion validity:		**fair**	
			Bouhassira (2004): Lower but acceptable correlations: pain with brushing (ρ= 0:70), pain due to pressure (ρ= 0.73); and pain due to cold (ρ= 0.66)			
0-10 NRS	xx	xx	xx	xx	***Responsiveness:*** *Farrar (2010):* On ROC analysis a raw change of -1.74 and a % change of -27.9% were associated with clinically meaningful change	**fair**
NTSS-6	***Internal consistency:*** *Bastyr (2005):* NTSS-6 showed satisfactory unidimensionality (Moderate, α= 0.7)	**poor**	***Construct validity:*** *Bastyr (2005):* NTSS-6 and NSC scores showed moderately positive and significant correlation. (ϒ= 0.773-0.885, p< 0.001)	**fair**	***Responsiveness: MCIDs*** *Bastyr (2005):* A change of 0.97 points showed a reasonable change for	**fair**

	test-retest reliability: *Bastyr (2005):* Satisfactory ICC scores with lower but acceptable test retest reliability (Baseline ICC= 0.900, End point ICC= 0.903)	**fair**	***Convergent validity:*** *Bastyr (2005):* NTSS-6 and NSC scores showed poorly positive and significant correlation with changes from baseline (ϒ= 0.519-0.708, p< 0.001)	**fair**	minimal improvement.	
0-10 point PI-NRS	xx	xx	xx	xx	***Responsiveness:*** *Farrar (2001):* On ROC analysis a raw change of -2, -2.5, and -3 were associated with least, average, and worst pains	**good**
PQAS	***Internal consistency:*** *Jensen (2010):* PQAS showed satisfactory unidimensionality: Deep scale (Moderate α= 0.75), surface scale (Poor α= 0.69), and paroxysmal scale (Good α= 0.87)	**fair**	***Construct validity:*** *Jensen (2010):* Three of the PQAS items and scale scores showed significant correlation with concurrent pain interference on BPI (p< .01)	**fair**	***Responsiveness:*** *Jensen (2006):* Ten of the PQAS descriptor items significantly picked up the changes in scores after treatment (p< .0025)	**poor**

P-NPSI	***test-retest reliability:*** *de Andrade (2011):* Satisfactory ICC scores with moderate test retest reliability (ICC= 0.7678)	**fair**	***Face validity:*** *de Andrade (2011):* P-NPSI was filled in less than 8 minutes by 85% of participants. Prevalence rate= 65% ***Construct validity:*** *de Andrade (2011):* PV-NSSI showed low but acceptable correlation with NRS: at first visit (Poor ρ= 0.40, p< 0.0001), at second visit (Poor ρ= 0.53, p< 0.0001), and change score (Poor ρ= 0.22, p< 0.0001)	**poor** **fair**	***Responsiveness:*** *de Andrade (2011):* PV-NPSI change scores show significant correlation with P-GIC (Good ρ= 0.727), and C-GIC scores (Poor ρ= 0.645)	**fair**
QST	***inter-rater reliability:*** *Geber (2011):* QST showed significant inter-rater reliability, r= 0.83 (range= 0.56- 0.89, p< 0.01) *Maser (1989):* 81% of inter-observer agreement that QST can be used adjacent to clinical examination for NeP assessment	**good** **fair**	***Construct validity:*** *Felix (2009):* QST showed significant correlation with average thermal pain threshold (r= 0.58 at p< 0.02)	**poor**	xx	xx

	test-retest reliability: *Felix (2009):* Low but acceptable ICC scores: cold, and hot pain (Poor ICCs= 0.50) *Geber (2011):* QST showed significant test-retest reliability, r= 0.86 (range= 0.67- 0.93, p< 0.01)	**poor** **good**				
SESWM	xx	xx	***Construct validity:*** *Schreuders (2008):* SESWM showed low but significant correlations with MMT (Poor r= 0.57), RIHM dynamometry (Poor r= 0.65, p< 0.001)	**fair**	xx	xx
SF-MPQ-2	***Internal consistency:*** *Dworkin (2009):* SF-MPQ-2 showed satisfactory unidimensionality: Web survey data (Excellent, α= 0.91), and clinical trial data (Excellent, α= 0.95)	**excellent**	***Construct validity:*** *Dworkin (2009):* SF-MPQ-2 scores showed significant correlation with rating of pain and sleep interference, BPI interference scale sores, the SF- 36 PCS, MCS scores, the HADS anxiety and depression subscale	**fair**	***Responsiveness:*** *Dworkin (2009):* Both total and sub-scale scores were responsive to changes that were meaningful to	**fair**

			scores		patients		
Spanish NPSI	***Internal consistency:*** *Villoria (2011):* S-NPSI showed satisfactory unidimensionality: total NPSI score (α> 0.80), and NPSI sub scores (α> 0.70)	**excellent**	***Construct validity:*** *Villoria (2011):* S-NPSI showed acceptable accuracy to detect responses of pain as defined by either the clinical or the discriminant criteria	**fair**	xx	xx	
	test-retest reliability: *Villoria (2011):* Moderate test-retest reliability with satisfactory ICC scores (0.680- 0.810)	**good**					
TCSS	***Inter-rater reliability:*** *Bril (2002):* Low but acceptable inter-rater reliability (6.3%)	**poor**	***Construct validity:*** *Bril (2002):* TCSS showed poor and inverse correlation with SUMAMP and SUMCV (ϒ= 0.424; ϒ= 0.302 at p< 0.0001; and p= 0.0044)	**good**	xx	xx	
	Intra-rater reliability: *Bril (2002):* Moderate and satisfactory intra-rater reliability (7.3%)	**poor**					

TNS	***inter-rater reliability:*** *Cornblath (1999):* Satisfactory ICC scores with excellent inter-rater reliability (ICC= 0.938, 95% CIs, p≥ 0.836)	**fair**	***Construct validity:*** *Cornblath (1999):* TNS showed significantly high and positive correlation with NIS (Good, ρ= 0.89, 95 % CIs) & NSS (Good, ρ= 0.86, 95% CIs)	**fair**	xx	xx
	intra-rater reliability: *Cornblath (1999):* Satisfactory ICC scores with excellent intra-rater reliability (ICC= 0.973, 95% CIs, p≥ 0.950)	**fair**				
TRNDSI	***test-retest reliability:*** *Collins (2008):* Satisfactory test-retest reliability for CRPS-I and Fibromyalgia (Excellent and Good, ICC= 0.93; and 0.83)	**good**	xx	xx	xx	xx
	Measurement error: *Collins (2008):* SEM values were small	**good**				

	compared with domain sum scores (3.5%- 8.3%)					
UENS	***inter-rater reliability:*** *Singleton (2008):* UENS showed a satisfactory high inter-rater reliability (94%)	**poor**	***Criterion validity:*** *Singleton (2008):* UENS (baseline and changeover scores) showed a close correlation with Michigan Diabetic Neuropathic scale and Neuropathy Impairment Score- Lower Leg ($p< 0.001$)	**excellent**	***Responsiveness:*** *Singleton (2008):* UENS showed a Good diagnostic sensitivity at baseline without sacrificing specificity	**poor**
VAS	xx	xx	***Hypothesis testing:*** *Davidoff (1988):* The VAS had significant correlations with limb volume ($r^2= 0.160$), active ROM (upper extremity: $r^2= 0.167$; lower extremity: $r^2= 0.508$)and joint pain ($r^2= 0.341$)	**poor**	xx	xx
ZBPI	***test-retest reliability:*** *Coplan (2004):* ZBPI showed low but acceptable test-retest reliability (Poor, ICC= 0.63 b/w 5-7 days; Moderate, ICC=	**good**	***Hypothesis testing:*** *Coplan (2004):* ZBPI showed satisfactory and acceptable correlations with MPQ (24 hours: $\gamma> 0.79$ and for 14-35 days $\gamma> 0.65$), ADL (for 14-35 days: γ	**good**	xx	xx

	0.78 b/w 8-10 days and 11-14 days after rash onset)	>0.52), and QoL (γ= 0.78) *Schmader (2007):* ZBPI showed a significant correlation with other domains. Increased composite pain and discomfort intensity scores were associated with increase in ZBPI ADL interference	**fair**

Abbreviations: ADL= Activities of Daily Living, BPI= Brief Pain Inventory Scale, BPI- DPN= Brief Pain Inventory Scale for Diabetic Peripheral Neuropathy, CGIC= Clinical Global Impression of Change, CPD= Chronic pain descriptors, CRPS= Complex Regional Pain Syndrome severity, DN4= Douleur Neuropathique 4, dSCT-2= diabetes Symptom Checklist Type-2, FFI= Foot Function Index, HADS= Hospital Anxiety and Depression Scale, QoL= Quality of Life, LANSS= Leeds Assessment of Neuropathic pain Symptoms and signs Screening Tool, MCS= Mental Component Summary, MPQ= McGill Pain Questionnaire, mTCNS= modified Toronto Clinical Neuropathy Score, NIS= Neuropathy Impairment Score, NPS= The Neuropathic Pain Scale, NPSI= Neuropathic Pain Sensory Inventory, NRS= Numeric Rating Scale, NSC= Neuropathy Symptom and Change score, NSS= neuropathy sensory symptoms, NTSS-6= Neuropathy Total Symptom Score-6, PGIC= Patient Global Impression of Change, PI-NRS= Pain Intensity Numeric Rating Scale, P-NPSI= Portuguese version of the Neuropathic Pain Symptoms Inventory, QST= Quantitative Sensory Testing, RIHM= Rotterdam Intrinsic Hand Myometer, SEM= Standard Error of Mean, SESWM= Sensory evaluation with Semmens-Weinstein Monofilaments, SF-MPQ= Short-form McGill Pain Questionnaire, SF-12= The Medical Outcomes Study Short Form Health Survey (SF-12), SUMAMP= Sum of lower limb distal amplitude, TCSS= Toronto clinical scoring system, TNS= Total Neuropathy Score, TRNDSI= The Trauma Related Neuronal Dysfunction Symptoms Inventory, UENS= The Utah Early Neuropathy Scale, VAS= Visual Analog Scale, VRS= Verbal Rating Scale, xx= not determined, ZBPI= Zoster Brief Pain Inventory

Table V Summary of identified physical functioning outcome measures with their published psychometric properties and COSMIN grading

OMs	Reliability	**COSMIN**	Validity	**COSMIN**	Responsiveness	**COSMIN**
AMHFQ	***Internal consistency:*** *Alderson (1999):* AMHFQ showed satisfactory unidimensionality (Excellent, α= 0.97) ***test- retest reliability:*** *Alderson (1999):* All the items showed consistent results with in 95th percentile confidence limits (Poor – Moderate ICCs)	**poor** **poor**	***Convergent validity:*** *Alderson (1999):* Poor correlation with dynamic two-point discrimination (ϒ= -0.32), static two-point discrimination (ϒ= -0.127), the Valpar upper extremity range of motion (ϒ= -0.2388), Pain VAS (ϒ= 0,36), functional VAS (ϒ= 0.3688), grip strength (ϒ= 0.3867), three point pinch strength (ϒ= 0.295), and lateral pinch strength (ϒ= 0.151)	**poor**	xx	xx
BI	xx	xx	***Construct validity:*** *Padua (2008):* Significant relationship b/w ability to walk on toes, strength of lower limbs muscles, abnormal stand-up, abnormal Romberg test, tactile sensory tests; medium relationship with ability to stand up and strength forearm and intrinsic hand	**fair**	xx	xx

				muscles; and lowest relationship with strength of hand intrinsic muscles			
BBM	xx	xx	***Hypothesis testing:*** *Erdmann (2005):* High BBS showed low correlation with 10 MWT and SIP68 scores (ρ= -0.76, and ρ= -0.62)	**fair**	xx	xx	
BPI-Facial	***Internal consistency:*** *Lee (2010):* BPI-Facial showed satisfactory unidimensionality: entire instrument (Excellent α= 0.94), intensity of pain (Good α= 0.86), interference with general activities (Good α= 0.89), and interference of facial- specific items (Excellent α= 0.95)	**fair**	***Construct validity:*** *Lee (2010):* BPI-Facial showed borderline significant correlation with NRS: At least amount of pain (1.01, p= 0.111), and during the week (0.95, p= 0.101)	**fair**	xx	xx	
CMTNS	***inter-rater reliability:*** *Shy (2005):* Satisfactory ICC scores with excellent inter-rater reliability	**fair**	***Construct validity:*** *Shy (2005):* CMTNS showed strong and satisfactory correlations with Ambulation Index (r= 0.81), Self-	**fair**	***Responsiveness:*** *Shy (2008):* CMTNS can be used satisfactorily to detect	***poor***	

	(ICC= 0.98, p<0.01)		Assessment Questionnaire (r= 0.76), Hand Function (r= 0.66), 9 Hole Peg test (r= 0.65), CMTNS ulnar and median CMAP amplitudes (r= 0.76, 0.72) and Neuropathy Impairment Score (r= 0.96)		progression of CMT disease	
	intra-rater reliability: *Shy (2005):* The scores from intra-scoring examination did not significantly vary on sensory evaluation	**fair**				
CMTNS-2	***inter-rater reliability:*** *Murphy (2011):* Satisfactory ICC scores with excellent inter-rater reliability: CMTSS2 (ICC= 0.97), and CMTES2 (ICC= 0.96)	**poor**	xx	xx	xx	xx
	intra-rater reliability: *Murphy (2011):* Satisfactory ICC scores with excellent intra-rater reliability: CMTSS2 (ICC= 0.96), and CMTES2 (ICC= 0.97)	**poor**				

DASH	***Internal consistency:*** *Dias (2008):* DASH showed satisfactory unidimensionality (Excellent, α= 0.98) *Novak (2010):* DASH showed satisfactory unidimensionality (Excellent, α= 0.96) ***test-retest reliability:*** *Dias (2008):* Lower test retest reliability (test-retest differences= -4.7 to 4.9, 95% CIs, p= 0.02	**poor** **poor** **poor**	***Construct validity:*** *Dias (2008):* DASH showed no significant correlations with Gartland and Worley scores (Y= -0.33, 5% level) *Zimmerman (2009):* DASH showed a significant correlation with grip strength (r= -0.53), and pinch strength (r= -0.49) *Novak (2010):* DASH showed a positive correlation with VAS for pain (Poor, r= 0.51, p< 0.001) ***Criterion validity:*** *Zimmerman (2009): DASH scores corresponded strongly with clinical staging (p< 0.001)* ***Hypothesis testing:*** *Eklund (2009):* DASH showed strong relationship b/w reduced hand function and upper-limb disability: manual dexterity (r= -0.64), finger dexterity (r= 0.83), grip strength (r= -0.72), tactile gnosis (r= -0.79), and hand function index (r= -0.71)	**poor** **fair** **fair** **poor**	xx		xx
DI	xx	xx	***Construct validity:***	**fair**	xx		xx

			Padua (2008): DI showed a significant relationship b/w ability to walk on toes, strength of lower limbs muscles, abnormal stand-up, abnormal Romberg test, tactile sensory tests; medium relationship with ability to stand up and strength forearm and intrinsic hand muscles; and lowest relationship with strength of hand intrinsic muscles				
DMM PUT	***test-retest reliability:*** *Amirjani (2011):* Satisfactory ICC scores with excellent test retest reliability (ICC= 0.91 at 95% CI, p< 0.001)	**fair**	***Hypothesis testing: Discriminative validity-*** *Amirjani (2011):* DMMPUT was significantly able to differentiate between impaired hand functions with mild, moderate and severe CTS	**fair**	xx		xx
FDI	***Internal consistency:*** *VanSwearingen (1996):* FDI showed a satisfactory unidimensionality (Theta reliability= 0.88)	**fair**	***Construct validity:*** *VanSwearingen (1996):* FDI physical function subscale showed a good correlation with clinician's physical examination of facial movements	**fair**	xx		xx
FDT	***test-retest reliability:*** *Videler (2008):* Satisfactory ICC scores	**fair**	xx	xx	xx		xx

	with good test retest reliability (ICC= 0.83-0.95, 95% CIs)						
HAP	***Internal consistency:*** *Farrell (1996):* HAP showed satisfactory unidimensionality (Excellent to Moderate α= 0.73- 0.97)	**poor**	***Hypothesis testing:*** *Farrell (1996):* HAP showed strong relationship with both maximum activity score and adjusted activity score (Excellent, r= 0.97, p< 0.000)	**fair**	***Responsiveness:*** *Farrell (1996):* HAP was sensitive enough to pick up changes in initial scores at the time of discharge	**poor**	
			Criterion validity: *Farrell (1996):* HAP showed strong correlation with maximum activity score (Good r= 0.78, p< .000), adjusted activity score (Good r= 0.83, p< 0.000), and Barthel Index: Self-care (Moderate r= 0.75, p< 0.000), mobilising (Poor r= 0.61, p< 0.000)	**good**			
INCAT ODSS	xx	xx	***Concurrent validity:*** *Merkies (2006):* INCAT ODSS showed low but significant association with changes in ODSS (Poor r= 0.66, p= 0.007), Rankin changes (Poor r=0.60, p=0.02), and GBS Disability Scale changes (Poor r= 0.56, p= 0.04)	**poor**	xx	xx	

ISS	***Internal consistency:*** *Merkies (2000):* ISS showed satisfactory unidimensionality: First visit (Poor α= 0.68), second visit (Moderate α=0.73), third visit (Moderate α= 0.71), and longitudinal (Good α= 0.87)	**poor**	***Construct validity:*** *Merkies (2000):* ISS showed moderate correlations with the additional scales in the stable group (Poor, r= 0.38- 0.56, p< 0.006)	**fair**	***Responsiveness:*** *Merkies (2000):* ISS showed significant association of patient's grading with the clinical judgment scores during follow up (p< 0.0001)	**poor**
	inter-rater reliability: *Merkies (2000):* Satisfactory ICC scores with good inter-rater reliability (ICC= 0.85 to 0.89, p< 0.0001)	**fair**				
	intra-rater reliability: *Merkies (2000):* Satisfactory ICC scores with good intra-rater reliability (ICC= 0.85 to 0.89, p< 0.0001)	**fair**				
LKQ	xx	xx	***Criterion validity:*** *Zimmerman (2009):* LKQ showed a significant correlation with DASH: symptom score (r= 0.79), and	**poor**	xx	xx

			function score (r= 0.87, p< 0.001)			
			Construct validity:	**poor**		
			Zimmerman (2009): LKQ function and symptom scores corresponded strongly with clinical staging (p< 0.001)			
MHQ	***Internal consistency:***	**poor**	***Construct validity:***	**poor**	xx	xx
	Dias (2008): MHQ showed satisfactory unidimensionality (Excellent, α= 0.93)		*Dias (2008)*: MHQ showed no significant correlations with Gartland and Worley scores (ϒ= -0.30, 5% level)			
	test-retest reliability:	**poor**				
	Dias (2008): Lower test retest reliability (test-retest differences= -4.3 to 2.2, 95% CIs, p= 0.02)					
mNDS	xx	xx	***Criterion validity:***	**good**	xx	xx
			Asad (2010): mNDS proved 92.31% sensitivity and 47% specificity in assessing the sensorimotor neuropathy			
10-MWT	***inter-rater reliability:***	**fair**	***Hypothesis testing:***	**poor**	xx	xx
	Solari (2008): Satisfactory inter-rater reliability with ICC= 0.97 (CI= 0.88-0.99)		*Erdmann (2005):* High 10 MWT scores correlated significantly with high SIP68 scores (ρ= 0.59, p= 0.036)			

	intra-rater reliability: *Solari (2008):* Satisfactory intra-rater reliability with ICC= 0.96 (CI= 0.87-0.99)	**fair**				
NHPT	***inter-rater reliability:*** *Solari (2008):* Satisfactory inter-rater reliability with ICC= 0.95 (CI= 0.89-0.97)	**fair**	xx	xx	xx	xx
	intra-rater reliability: *Solari (2008):* Satisfactory intra-rater reliability with ICC= 0.95 (CI= 0.89-0.97)	**fair**				
NIS	xx	xx	xx	xx	***Responsiveness***: *Shy (2008):* NIS can be used satisfactorily to detect progression of CMT disease	**poor**

ODSS	***inter-rater reliability:*** *Merkies (2002)*: Satisfactory ICC scores with excellent inter-rater reliability: Experienced examiners (ICC= 0.95), Variable examiners (ICC= 0.90)	**fair**	***Construct validity:*** *Merkies (2002):* ODSS showed low correlation with MRC (Poor r= 0.45), INCAT sensory sum score (Poor r= 0.41), and Right & left hand grip strengths (Poor r= 0.54 & 0.53)	**fair**	***Responsiveness:*** *Merkies (2002):* Scores showed significant association with clinical changes during follow ups (Poor r= 0.66, p= 0.008)	**good**
	intra-rater reliability: *Merkies (2002):* Satisfactory ICC scores with excellent intra-rater reliability: Experienced examiners (ICC= 0.95), Variable examiners (ICC= 0.93)	**fair**				
ONLS	***Internal consistency:*** *Graham (2006):* ONLS showed satisfactory unidimensionality (Poor, α= 0.6)	**fair**	***Content validity:*** *Graham (2006):* The results showed that ONLS is appropriate to use in clinical practice	**fair**	***Responsiveness:*** *Graham (2006):* ONLS was capable enough to capture a change in activity measures to a similar extent as that of ODSS (SRM= 0.76, 95% CIs)	**poor**
	inter-rater reliability: *Graham (2006):* Satisfactory ICC scores with excellent test retest reliability (ICC= 0.97)	**poor**	***Construct validity:*** *Graham (2006):* ONLS showed a variable correlation with ODSS (Excellent, r= 0.97, p<0.001), 10-meter walk time (Poor, r= 0.58), and MRC score (Poor, r= -0.62)	**fair**		

	Solari (2008) Satisfactory inter-rater reliability with weighted kappa for arm score= 0.65 (95% CI= 0.44-0.86), and weighted kappa for leg score= 0.63 (95% CI= 0.41- 0.85)	**fair**					
	intra-rater reliability: *Solari (2008):* Satisfactory intra-rater reliability with weighted kappa for arm score= 0.75 (95% CI= 0.54-0.96), and weighted kappa for leg score= 0.68 (95% CI= 0.47- 0.90)	**fair**					
	test-retest reliability: *Graham (2006):* ONLS showed acceptable test-retest reliability as 15 neurologists independently preferred ONLS	**poor**					
PEM	***Internal consistency:*** *Dias (2008):* PEM showed satisfactory	**poor**	***Construct validity:*** *Dias (2008):* PEM showed no significant correlations	**poor**	xx		xx

	unidimensionality (Excellent, α= 0.94)		with Gartland and Worley scores (ϒ= -0.37, 5% level)				
	test-retest reliability:	**poor**					
	Dias (2008): Lower test retest reliability (test-retest differences= -9.3 to 2.3, 95% CIs, p= 0.02)						
PPMs	***test-retest reliability:*** *Manor (2008):* Both 6 minute walk test and Timed up and go test showed significant reliability (Excellent ICC= 0.93- 0.99, 95% CIs)	**poor**	xx	xx	xx		xx
QRS	**test-retest reliability:** *Perez (2002):* QRS showed satisfactory ICC scores with good test-retest reliability (range= 0.84- 0.87, p< 0.001)	**poor**	xx	xx	xx		xx
RSQ	***inter-rater reliability:*** *Oerlemans (2000):* For inter-rater reliability the limits of agreement between two observers was -0.26 and	**poor**	***Construct validity:*** *Oerlemans (2000):* For observer A, 11 test categories were highly correlated (> 0.80), however for observer B, the correlations were lower (but mostly > 0.60)	**poor**	xx		xx

	0.22 ***test-retest reliability:*** *Oerlemans (2000):* For test-retest reliability the limits of agreement between observer A (-0.10 and 0.14) and observer B (-0.26 and 0.22) was very close	**poor**				
SALSA	***Internal consistency:*** *The SALSA Collaborative Study Group (2007):* SALSA showed satisfactory unidimensionality: Leprosy group (Good, α= 0.897), and diabetes group (Good, α= 0.814)	**excellent**	***Construct validity:*** *Melchior (2011):* SALSA showed low but acceptable correlation with NPHT (Moderate r=0.77, p<.0005), SHFE (Poor r= 0.66, p<.0005), and FDT (Poor r= 0.54, p<.005) ***Content validity:*** *The SALSA Collaborative Study Group (2007):* SALSA showed strong relationship to the scores assigned by independent experts: Overall (ρ= 0.67), leprosy group (ρ= 0.65), and diabetes group (ρ= 0.70)	**poor** **excellent**	xx	xx
SAM	xx	xx	***Hypothesis testing:***	**poor**	xx	xx

				Smith (2004): SAM showed a strong correlation with physical Function scale, Physical Component Summary score, and Vitality scale (p= 0.01); and a weak correlation with Bodily Pain and Role Limitation (p= 0.05)			
SAM (4 mWT)	xx	xx		***Construct validity:*** *van Schie (2011):* SAM (4mWT) showed a significant correlation with Dutch version of International Physical Activity Questionnaire: min/ week (p= 0.49), and activity/ week (p= 0.43, p< 0.05)	**poor**	xx	xx
				Criterion validity: *van Schie (2011):* SAM recorded an accuracy of 98.6% compared with observer- counted strides	**poor**		
SDS	***Internal consistency:*** *Rejas (2008):* SDS showed satisfactory unidimensionality (Excellent, α= 0.904)	**poor**		xx	xx	**Responsiveness:** *Rejas (2008):* SDS was significantly able to differentiate between responders and non-	**fair**

					responders	
SHFT	***Internal consistency:*** *Videler (2008):* SHFT showed excellent homogeneity for both dominant hands (α= 0.96), and non-dominant hands (α= 0.95)	**fair**	xx	xx	xx	xx
	test-retest reliability: *Videler (2008):* SHFT showed satisfactory test-retest reliability with good ICC (83- 0.95, 95% CIs)	**fair**				
TBQ	***Internal consistency:*** *Sezgin (2006):* TBQ showed satisfactory unidimensionality: symptom severity scale (Good α= 0.82), and function status scale (Good α= 0.88)	**poor**	***Construct validity:*** *Sezgin (2006):* TBQ showed satisfactory correlations with symptoms severity scale (r= 0.73, p< 0.00001); moderate and good correlations with subscales of SF-36- physical functioning (r= 70.55), physical role (r= 70.54), bodily pain (r= 70.63, p< 0.0001), and emotional role (r= 70.40, p< 0.001)	**good**	xx	xx
	test-retest reliability: *Sezgin (2006):* TBQ showed	**fair**				

	satisfactory correlation scores with acceptable test-retest reliability: symptom severity scale (Poor, r= 0.60), and function status scale (Moderate r= 0.77, p= 0.0001)						
UNEQ	***Internal consistency:*** *Mondelli (2006):* UNEQ showed satisfactory unidimensionality (Good, α= 0.87)	**poor**	***Content validity:*** *Mondelli (2006):* UNEQ showed a satisfactory content validity as all the questions were equally distributed between the symptoms numbness/tingling and elbow pain	**excellent**	***Responsiveness:*** *Mondelli (2006):* UNEQ showed significant responsiveness in picking up difference in scores at follow ups (Good, r=0.85, p<0.001)	**poor**	
	test-retest reliability: *Mondelli (2006):* Satisfactory ICC scores with excellent test retest reliability (ICC= 0.97)	**fair**	***Construct validity:*** *Mondelli (2006):* UNEQ showed satisfactory correlations with scores of the clinical (Poor, ρ=0.65) and electrophysiological (Poor, ρ=0.35) severity scales	**poor**			
Walk-12	***Internal consistency:*** *Graham (2006):* Walk-12 showed satisfactory unidimensionality (Excellent, α= 0.97)	**poor**	***Hypothesis testing:*** *Graham (2006):* Walk-12 showed strong correlation with the SF-36 Physical Function Subscale (r= 20.82), the Social Function Component (r= 20.86), Physical	**poor**	xx	xx	

	test-retest reliability: *Graham (2006):* Satisfactory ICC scores with excellent test retest reliability (ICC= 0.96)	**poor**	Component Summary Score (r= 20.72) and the lower limb section of the ONLS (r= 0.77)		
WSQ	***test-retest reliability:*** *Perez (2002):* WSQ showed satisfactory ICC scores with moderate test-retest reliability (range= 0.78-0.87, p< 0.001)	**poor**	xx		xx
WST	***test-retest reliability:*** *Kilmer (2000):* WST showed acceptable test-retest reliability: Pronation (Good ICC= 0.88), supination (Good ICC= 0.85), push (Excellent ICC= 0.96), pull (Excellent ICC= 0.93), and lever arm push (Poor ICC= 0.67)	**poor**	***Construct validity:*** *Kilmer (2000):* WST showed strong and positive correlations with Hand Held Dynamometry- measured peak torque for both dominant and non-dominant hands (p< 0.05)	**poor**	xx

ZIQ	xx	xx	***Hypothesis testing:*** *Schmader (2007):* ZIQ showed a significant correlation with other domains. Increased composite pain and discomfort intensity scores were associated with increase in ZIQ ADL interference scores	**fair**	xx

Abbreviations: ADL= Activities of Daily Living, AMHFQ= The Alderson-McGall hand function questionnaire, BBM/S= Berg Balance Measure/ Score, BI= Barthel Index, BPI= Brief Pain Inventory, CMAP= compound muscle action potential, CMTNS= Charcot-Marie-Tooth disease Neuropathy score, DASH= The Disabilities of Arm, Shoulder and Hand Questionnaire, DI= Deambulation Index, DMMPUT= Dellon-modified Moberg pick-up test, FDI= Facial disability Index, FDT= Functional dexterity test, GBS= Guillain Barré Syndrome, HADS= Hospital Anxiety and Depression Scale, HAP= Human Activity Profile, ISS= Inflammatory neuropathy Sensory Score, LKQ= Levine-Katz Questionnaire, MHQ= The Michigan Hand Outcome Questionnaire, 10-MWT= 10-Meter walking test, mNDS= modified Neuropathy Disability Score, NHPT= Nine-Hole Peg test, NIS= Neuropathy Impairment Score, NRS= Numeric Rating Scale, ODSS= The Overall Disability Sum Score, ONLS= The Overall Neuropathy Limitations Scale, PEM= The Patient Evaluation Measure, PPMs= Physical Performance Measures (6 minute walk test, Timed up and go test), QRS= Questionnaire rising and sitting down, R36HS= Rand-36 Health Survey, RSQ= The Radboud skills Questionnaire, SALSA= Screening of Activity Limitation and Safety Awareness Scale, SAM= Step Activity Monitor, 4mWT= 4 min walk test, SDS= Sheehan Disability Scale, SHFT= Sollerman Hand function test, SIP68= Sickness impact profile 68, TBQ= Turkish version of the Boston Questionnaire, UNEQ= Ulnar neuropathy at the elbow Questionnaire, VAS= Visual Analog Scale, Walk-12= 12-Item Multiple Sclerosis Walking Scale, WSQ= Walking stairs Questionnaire, WST= Work stimulation tasks (knob turn, Linear motion, and Lever arm), xx= not determined, ZIQ= Zoster Impact Questionnaire

Table VI Definition of domains, measurement properties, aspects of measurement properties and accepted statistical analyses by COSMIN

Domain	Measurement property	Aspect of a measurement property	Definition	Accepted statistical analyses	Interpretation	Inappropriate statistical analyses
Reliability	Internal consistency		The degree of the interrelatedness among the items	Cronbach's alpha (α) Internal consistency coefficient	α> 0.90: Excellent α= 0.81- 0.90: Good α> 0.71-0.80: Moderate α< 0.70: Poor	Pearson's correlation coefficient Spearman's correlation coefficient
	Reliability	*Intra-rater reliability; Inter-rater reliability; test-retest reliability*	The proportion of the total variance in the measurements which is due to 'true' differences among patients	Continuous scores: ICC Dichotomous/nominal scores: Cohen's kappa (Κ) Ordinal scores: Weighted kappa	ICC or Κ> 0.90: Excellent ICC or Κ=0.81-0.90: Good ICC or Κ> 0.71-0.80: Moderate ICC or Κ< 0.70: Poor	
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured	SEM, SDC or LoA		
Validity	Content validity		The degree to which the content of a HR-PRO is an adequate reflection of the construct to be measured			
		Face validity	The degree to which (the items	Requires a subjective		

		of) an instrument indeed looks as though they are an adequate reflection of the construct to be measured	judgement, thus no analytical standards are developed	
Construct validity		The degree to which the scores of a HR-PRO are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured		
	Structural validity	The degree to which the scores of a HR-PRO are an adequate reflection of the dimensionality of the construct to be measured	Factor analysis	
	Hypotheses testing- Discriminant validity; Convergent validity; Divergent	Idem construct validity	Correlation coefficient	Positive correlation: γ> 0.90: Excellent γ= 0.81- 0.90: Good γ> 0.71-0.80: Moderate γ< 0.70: Poor Inverse correlation:

	validity; *Sensitivity & specificity*			γ< -0.90: Excellent γ= -0.81 to -0.90: Good γ= -0.71 to -0.80: Moderate γ>-0.70: Poor
	Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument	Confirmatory factor analyses Differential item functioning analyses	
Criterion validity	*Concurrent validity*	The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'	When both scores are continuous: Correlation co-efficient When one is continuous score and other is dichotomous: Area under the ROC When both scores are dichotomous: sensitivity & specificity	

Responsiveness	Responsiveness	The ability of an HR-PRO instrument to detect change over time in the construct to be measured	When both scores are continuous: Correlation co-efficient When one is continuous score and other is dichotomous: Area under the ROC When both scores are dichotomous: sensitivity & specificity	Effect size Standardised response mean Norman's responsiveness coefficient Relative efficacy statistic Guyyatt's responsiveness ratio MIC Paired t-test
Interpretability		The degree to which one can assign qualitative meaning- i.e., clinical or commonly understood connotations- to an instrument's quantitative scores or change in scores	MIC and MID	

Abbreviations: α= Cronbach's alpha, HR-PRO= Health related- patient reported outcome, ICC= Intra class correlation coefficient, K= Cohen's Kappa, LoA= Limits of Agreement, MIC= Minimal important change, MID= Minimal important difference, γ= Correlation coefficient, ROC= Receiver operating curve, SDC= Smallest Detectable Change, SEM= Standard Error of Measurement

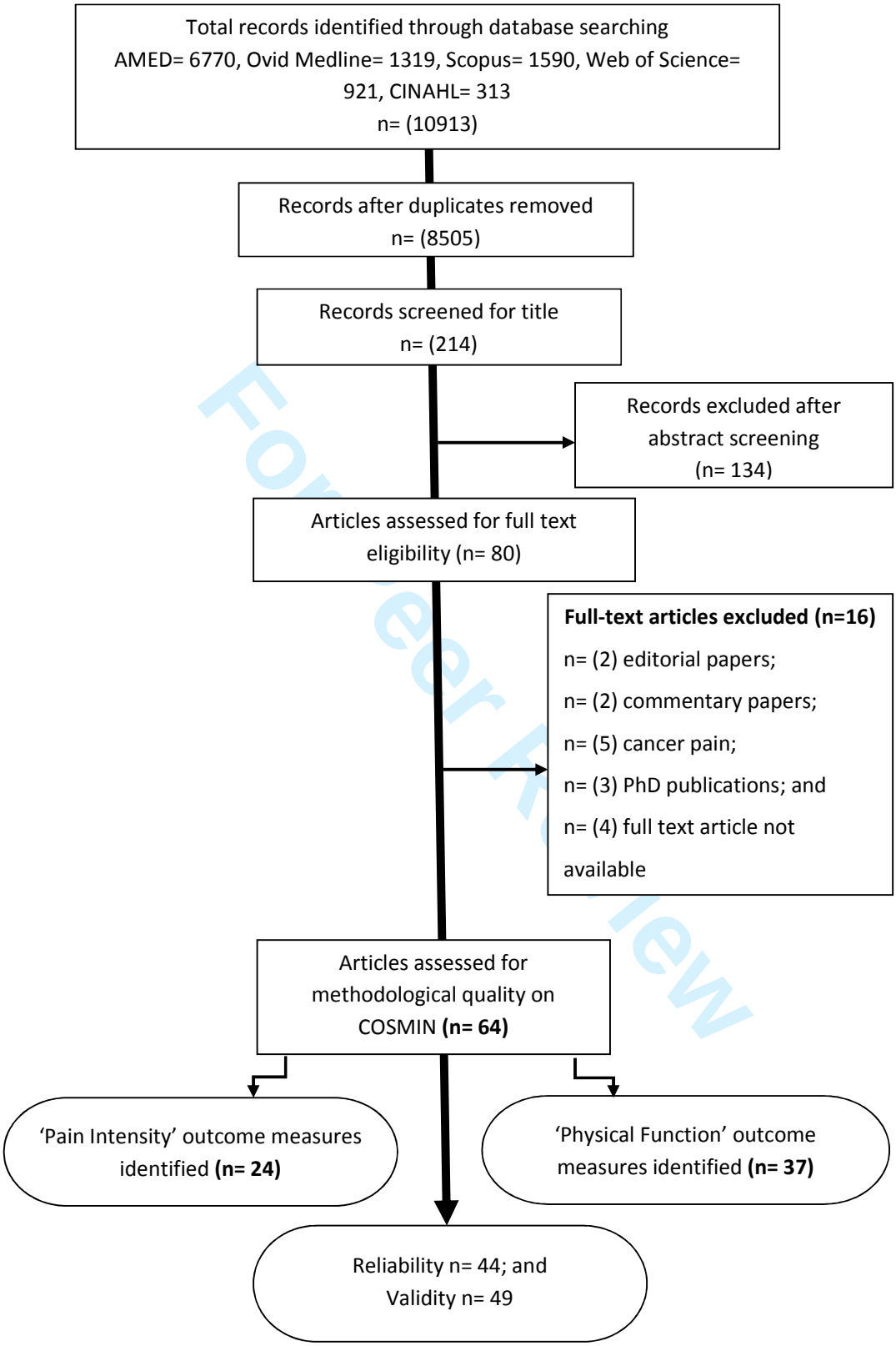

Figure I Flow diagram summarising study selection process

Reviewer 1		
COMMENT	EXPLAINATION	MODIFICATIONS (Highlighted text)
Introduction		
I think it is preferable to avoid using too many abbreviations e.g. PMP and OM.	We note the potential for confusion, thanks for this suggestion.	Necessary amendments on the specified pages have been made.
It would be helpful in the introduction to separate out the two concepts of 1) the need to test the psychometric properties of outcome measures-e.g. if reliability has been completed did the results indicate that the test is actually reliable and therefore could be recommended for use;[in methods would be good if you assessed this also i.e. quality of the results of measurement properties] from 2) the methods used to test the psychometric properties (e.g. with COSMIN). The objective gets lost within the final paragraph-can I suggest you rephrase as an aim and move the detail on COSMIN to your methods section.	Thank you for your comment. We note the reviewer's concern here. And hence the required explanation has been added as indicated.	Sentences explaining the aims and objectives of the study have been rephrased.
Method		
Page 4: Line 41 replace 'has also been activated' to 'was activated'; consider rephrasing this sentence as it is not very clear.	Agreed.	Corrections have been made in the text.
Check end search date-differs between abstract and methods.	Agreed.	Corrections have been made in the text.
Please clarify line 56 'OMs used in intervention trials….' with the statement on page 5-eligibility criteria which states that cross sectional clinical trials	We note the potential for confusion. The inclusion criteria for this study was the	The term unnecessary words have been deleted to avoid the confusion.

(what is this??, can you have a cross sectional intervention trial) and cohort studies (so do you mean an uncontrolled intervention study)?	cross sectional studies and the longitudinal cohort studies.	
Page 6: You seem to only describe a method to explore the methodological quality of the individual studies; there is no section on how you made a judgement on 'the evidence for the psychometric properties' as indicated in your objective on page 4; and there is no method section to describe how the results will be synthesised (so how can you temper the findings on reliability with the quality of the study-e.g. the study reports that the measure is very reliability but the methodological quality is very low).	Thanks for this comment. We concur with the reviewer's statement here.	The required explanation has been added under the section of data extraction and synthesis. A new table- Table VI has been added explaining the information of the criteria used for synthesizing the results of the study.
Results		
Page 8: It would be very helpful if you were able to add some description in the text to summarise the physical function outcomes measures-so were they self-report, physical performance, measuring ability e.g. steps versus disability. A similar overview of pain (if possible) would be helpful.	Agreed.	Please see manuscript for suggested overview.
Page 9: It would be important for the reader to know the results of the reliability tests as well as the methodological quality of the study which reported on these results (this would help inform some of the statements in your discussion e.g. page 10, line 53-many OMs seem promising'-on what basis?). So which tests were reliable (need to indicate in your methods how you made that judgement).	Thanks for this suggestion. We concur with reviewer's concern here.	Table VI has been added to the manuscript, explaining about the judgement criteria used for the studies.
Line 31-35-can you provide evidence to support your statement that 'these measures have been proven for their PMPs'.	We note the reviewer's concern here.	Reference has been provided in the text along with Table VI.

Page 9: It would be helpful to describe in a separate section the results for each of COSMIN boxes that you used.	Considering the magnitude of the COSMIN (9 boxes of definitions and explanation for each psychometric property for each outcome measure) and the word limit, explaining about the results of the studies in the form of paragraph seemed to a mere replication of the tables and thus was avoided.	No modifications made.
Discussion		
I found the discussion challenging to read as the text of the results did not present the results of the psychometric property under test e.g. if reliability was being tested was many of the tests were reliable-and then tempering these findings by only using results from the higher quality studies-you may have done this but it is not explicit to me in your reporting. I think the discussion would become more focused if the methods and results were expanded as I have suggested.	We concur with the reviewer's statement here. But considering the word count, explaining about the results of the studies in discussion seemed to a mere replication of the tables and thus was avoided. However the important facts which lead to the results and needs to be highlighted are well explained.	Necessary modifications have been made. The suggestion under the methods and results sections have also been accepted.
Reviewer 2		
COMMENT	EXPLAINATION	MODIFICATIONS (Highlighted text)
Well done. I have annotated the PDF with some minor grammatical errors; otherwise, the manuscript is well done.	Thanks for your feedback. The potential grammatical mistakes have been	Necessary modifications have been made in the sections of Abstract,

	corrected as per your advice.	Introduction, Methodology, Results, and Discussion.