INSTITUTO POLITÉCNICO DE LEIRIA

ESCOLA SUPERIOR DE SAÚDE

CURSO DE LICENCIATURA EM FISIOTERAPIA - TL4

Qualidade dos Instrumentos de Autorresposta que medem a Funcionalidade do Membro Superior em Condições Músculo-Esqueléticas do Ombro – Revisão Sistemática

Autores:

Ana Carolina Ferreira Freitas

Carina das Neves Vieira

Jéssica Elizabete Silveira

Joana Patrícia Silva Palaio

Leiria, Junho de 2016.

INSTITUTO POLITÉCNICO DE LEIRIA

ESCOLA SUPERIOR DE SAÚDE

CURSO DE LICENCIATURA EM FISIOTERAPIA - TL4

Monografia: Qualidade dos Instrumentos de Autorresposta que medem a Funcionalidade do Membro Superior em Condições Músculo-Esqueléticas do Ombro – Revisão Sistemática

Objetivos: Realizar uma revisão sistemática acerca da qualidade dos instrumentos de medida na funcionalidade do Complexo Articular do Ombro, traduzidos e validados para a população portuguesa.

Autores:

Ana Carolina Ferreira Freitas nº5120200

Carina das Neves Vieira nº5120210

Jéssica Elizabete Silveira nº5120208

Joana Patrícia Silva Palaio nº5120213

Orientador: Nuno Morais

Unidade Curricular: Monografia

Docentes: José Alves-Guerreiro, Luís Carrão, Nuno Morais, Sandra Amado

Leiria, Junho de 2016.

AGRADECIMENTOS

Considerando esta monografia o resultado de todo o percurso efetuado durante estes quatro anos de formação, agradecemos de antemão a todos os que, de alguma forma, marcaram o nosso percurso, contribuindo para o nosso crescimento quer a nível profissional quer pessoal. De forma particular, agradecemos:

- Aos nossos familiares, que não mediram esforços para que chegássemos ao final desta etapa, estando sempre presentes com o seu carinho, apoio e motivação.

- Ao professor orientador Nuno Morais que nos guiou ao longo desta caminhada e nos inspirou sempre a irmos mais além, ultrapassando os limites da nossa ambição.

- A todos os colegas de curso que, de alguma maneira, deram apoio ao longo de todo o processo de concretização deste trabalho.

A todos, um grande obrigado por estarem presentes
nesta enriquecedora experiência.

LISTA DE ABREVIATURAS

AAC – Área Abaixo da Curva ROC

APED – Associação Portuguesa para o Estudo da Dor

CAO – Complexo Articular do Ombro

CCI – Coeficiente de Correlação Intraclasse

COSMIN –*COnsesus-based Standards for the selection of health Measurement INstruments*

DASH- *Disabilities of the Arm Shoulder and Hand*

EPM – Erro Padrão de Medição

LC – Limites de Concordância

MAD – Menor Alteração Detetável

MAI – Mínima Alteração Importante

NULI-20- *Neck and Upper Limb Index*

PROM – *Patient Report Outcome Measures*

RCAAP – Repositório Científico de Acesso Aberto de Portugal

SF-12- *Short-form Health Survey 12-Item*

SF-36- *Short-form Health Survey 36-Item*

SPADI- *Shoulder Pain And Disability Index*

SRQ-PT- *Shoulder Rating Questionnaire*

UEFI- *Upper Extremity Functional Index*

WUSPI- *Weelchair User's Shoulder Pain Index*

RESUMO

Contextualização: Dada a prevalência de disfunções no ombro, os registos de avaliação segundo a perspetiva do utente constituem ferramentas úteis na seleção das estratégias de intervenção. A escolha do instrumento adequado deve-se basear em grande parte na força das suas propriedades psicométricas, contudo não existem estudos que analisem sistematicamente a qualidade destas medidas.

Objetivo: Análise de estudos referente às propriedades psicométricas de instrumentos de autorresposta na funcionalidade do ombro.

Metodologia: Revisão da literatura em inglês/português, nas bases de dados: PubMed, PEDro, Google Académico, B-On e RCAAP. Foram analisados estudos realizados até 2015. A qualidade metodológica e as propriedades psicométricas foram avaliadas e resumidas através de dois critérios padronizados, seguindo a ideologia COSMIN.

Resultados: Nesta revisão foram incluídos 6 estudos. O *Disabilities of the Arm Shoulder and Hand* (DASH) e o *Neck and Upper Limb Index* (NULI-20) demonstram boas propriedades psicométricas e uma metodologia de fraca a excelente; o *Shoulder Pain And Disability Index* (SPADI) e o *Weelchair User's Shoulder Pain Index* (WUSPI) exibem boas propriedades psicométricas e qualidade metodológica fraca; tanto no *Shoulder Rating Questionnaire* (SRQ-PT) como no *Upper Extremity Functional Index* (UEFI) não foram avaliadas propriedades psicométricas relevantes, contudo as analisadas apresentam boas propriedades psicométricas e uma metodologia fraca.

Conclusão: Devido às falhas na metodologia dos estudos incluídos, não é possível inferir qual o questionário mais apropriado à prática clínica. São necessários mais estudos de validação de instrumentos de autorresposta com melhor qualidade metodológica.

Palavras-chave: Ombro, Propriedades Psicométricas, Instrumentos de Autorresposta, COSMIN.

ABSTRACT

Background: Given the prevalence of shoulder dysfunctions, patient reported outcome measures (PROM) are useful tools for choosing intervention strategies. The instrument must be selected mainly according the strength of its psychometric properties. However, there aren't any studies that systematically analyze the quality of these measures.

Purpose: To analyze studies referring the psychometric properties of PROM in the shoulder.

Methods: Literature searches were performed in english/portuguese languages in the following databases: PubMed, PEDro, Google Scholar, B-On e RCAAP. We analyzed studies produced until 2015 inclusive. The methodological quality and its psychometric properties were accessed and summarized through two standardized criteria, following the COSMIN ideology.

Results: In this review they were included 6 studies. Disabilities of the Arm Shoulder and Hand (DASH) and Neck and Upper Limb Index (NULI-20) show good psychometric properties and weak to excellent methodology; both Shoulder Pain And Disability Index (SPADI) and Weelchair User's Shoulder Pain Index (WUSPI) exhibit good psychometric properties but poor methodological quality; both Shoulder Rating Questionnaire (SRQ-PT) and Upper Extremity Functional Index (UEFI) relevant psychometric properties have not been evaluated, however the analyzed ones have good psychometric properties and poor methodology.

Conclusions: Because of flaws in the methodology of the included studies, it is not possible to infer the most appropriate questionnaire in clinical practice. More studies are needed for the validation of PROM with better methodological quality.

Keywords: Shoulder, psychometric properties, patient reported outcome measures, COSMIN.

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE IMAGENS

1.INTRODUÇÃO

No âmbito da Unidade Curricular de Investigação Aplicada, inserida no 4º ano do Curso de Licenciatura em Fisioterapia da Escola Superior de Saúde, pertencente ao Instituto Politécnico de Leiria, foi proposto aos discentes a realização de um trabalho final de curso inédito em Portugal na área da saúde. Este trabalho tem como principal objetivo dar continuidade ao projeto selecionado previamente, apresentado na Unidade Curricular Investigação Aplicada.

A temática aqui presente relaciona-se com o estudo das propriedades psicométricas dos instrumentos de autorresposta que avaliam a funcionalidade do ombro, devidamente validadas para a população portuguesa, tendo como intuito realizar uma compilação deste tipo de instrumentos de avaliação por meio de uma revisão sistemática. As revisões sistemáticas apresentam um alto nível de evidência científica, uma vez que demonstram um especial cuidado em selecionar estudos pertinentes publicados e não publicados e a capacidade de avaliar os mesmos, sintetizando a informação de uma forma equilibrada e imparcial (Davies & Crombie, 2001). Aquando da realização de um estudo deste tipo, importa considerar cinco pontos fundamentais: a formulação de uma questão clara e concisa, a identificação dos estudos já realizados e que se manifestem relevantes, a avaliação da qualidade dos estudos, o resumo da evidência recolhida e a interpretação dos resultados (Khan, Kunz, Kleijnen, & Antes, 2003).

Desta forma, este estudo inicia-se com o enquadramento do problema, abrangendo a pertinência do tema (dados epidemiológicos e implicações na vida dos indivíduos portadores de patologia ao nível do ombro), a importância de uma boa avaliação no tratamento destas disfunções, quais os tipos de instrumentos de medida mais adequados e apresentação dos conceitos gerais das propriedades psicométricas. Neste documento, encontra-se descrito todo o processo subjacente à metodologia, nomeadamente os métodos de pesquisa, critérios de seleção, instrumentos de avaliação da qualidade metodológica dos estudos e de avaliação das propriedades psicométricas e métodos de síntese. Posteriormente, são apresentados os resultados obtidos bem como a discussão e conclusões retiradas dos mesmos.

2.FUNDAMENTAÇÃO

A dor no ombro é tipicamente caracterizada por sintomas nas várias articulações, músculos, tendões e bursas envolvidos no movimento do Complexo Articular do Ombro (CAO). O aparecimento de dor no ombro é variável e pode ocorrer sem causa direta ou estar relacionado com traumas, movimentos repetitivos ou eventos neurológicos (APED, 2010a).

Uma revisão sistemática que reúne dados epidemiológicos acerca da dor no ombro na população mundial, analisou dezoito estudos sobre prevalência e apenas um referente à incidência. Relativamente a esta taxa, a percentagem de dor no ombro varia entre 0.9 e 2.5% para diferentes idades. No que toca à prevalência, os dados diferem bastante, sendo que, de toda a população que sofre com esta problemática 6.9% a 26% referem-se a situações pontuais, 18.6% a 31% para intervalos de um mês de sintomatologia, 4.7% a 46.7% para um ano de presença de sintomas e 6.7% a 66.7% dizem respeito a situações álgicas que se repetem há mais de um ano (Luime et al., 2004).

Em Portugal, segundo dados da Associação Portuguesa para o Estudo da Dor (APED), a dor no ombro é a mais comum depois da dor na região lombar e no joelho. No período de um ano a prevalência total varia de 14% para 21%. Dentro de toda a população que padece de dor músculo-esquelética, 18% dos pagamentos de seguro de invalidez remetem para utentes com distúrbios cervicais e do ombro (APED, 2010a). Segundo outro estudo da APED, verifica-se uma prevalência para dor cervical e do ombro de 15% a 20%, sendo que esta é cerca de 1.5 vezes mais comum em mulheres do que em homens (APED, 2010b). Uma das patologias mais frequentes a nível nacional é a tendinopatia da coifa dos rotadores como resultado da realização de atividades que exigem a elevação mantida ou repetida dos membros superiores ao nível dos ombros ou acima destes ou ainda da realização de movimentos de circundação com os braços elevados durante a atividade laboral (DGS, 2008).

Os distúrbios do CAO são uma razão comum para as pessoas procurarem cuidados de saúde, com uma incidência anual estimada de 12/1000 consultas a gabinetes médicos.

Este tipo de patologias é de difícil resolução, já que apenas 50% dos pacientes com novos episódios de transtorno no ombro experienciam recuperação completa em seis meses, sendo que esta taxa aumenta para 60% após um ano. Os tratamentos mais comuns para as doenças do ombro incluem injeções de corticosteroides, manipulação articular, Fisioterapia e cirurgia, sem vantagem evidente de um tratamento sobre outro (APED, 2010a).

Dada a relevância deste problema de saúde pública, importa conhecer as implicações funcionais subjacentes ao mesmo. O CAO tem um papel fundamental na funcionalidade do utente uma vez que, devido à biomecânica do membro superior, esta articulação é a que confere função ao braço e mão (Paternostro-Sluga & Zöch, 2004). Os utentes com patologias no CAO podem apresentar quadros álgicos, restrição na amplitude de movimento ou diminuição da força muscular. A existência de dor ou lesões no ombro, para além do comprometimento a nível estrutural e funcional, podem afetar significativamente a situação social e profissional dos indivíduos (Paternostro-Sluga & Zöch, 2004). Deste modo os indivíduos com este tipo de condição terão dificuldades ou serão incapazes de executar várias atividades da vida diária, tais como: autocuidado, tarefas domésticas que exijam levantar e transportar objetos elevando o membro superior acima do nível do ombro e realizar movimentos amplos para o lado e para trás. Poderão ainda sofrer perturbações do sono, irritabilidade e alterações de humor (Winter, Heijden, Scholten, Windt, & Bouter, 2007).

Desta forma o tratamento de lesões referentes ao ombro é bem-sucedido se, para além de melhorar os problemas estruturais do indivíduo, recuperar também as suas atividades e participação na sociedade (Paternostro-Sluga & Zöch, 2004). Para tal é primordial a correção de problemas como a dor, restrição de mobilidade, falta de coordenação, padrões de movimentos compensatórios e fraqueza muscular (Hanratty et al., 2012; Paternostro-Sluga & Zöch, 2004). Portanto, uma intervenção fundamentada e informada é parte crucial do processo de tratamento.

Atualmente, a prática baseada na evidência é o elemento central do exercício clínico. Esta é fundamental para que os utentes recebam tratamentos eficazes, assim como para a redução dos custos em saúde. Para uma melhor tomada de decisão, para além da evidência científica, é fulcral uma avaliação rigorosa do utente, bem como a elaboração

de registos acerca do mesmo. Posto isto, temos os registos de avaliação como ferramentas extraordinariamente úteis na informação sobre a efetividade das intervenções e, consequentemente, para a continuidade das estratégias de intervenção utilizadas ou a sua alteração. Para além de permitirem aferir a efetividade da intervenção, constituem uma mais-valia na comunicação entre os diversos profissionais de saúde em equipas multidisciplinares, permitindo uma abordagem holística do indivíduo sem perdas de informação acerca da verdadeira condição do mesmo. Permitem ainda monitorizar os progressos e demonstrar a eficácia das intervenções aos utentes, instituições ou sociedade em geral (Hatfield & Ogles, 2007; Silva, 2006).

A demonstração da efetividade resulta não só da avaliação de parâmetros clínicos mas também de critérios que avaliem o impacto dessas alterações na vida dos utentes, na sua funcionalidade e qualidade de vida. Estes oferecem uma perspetiva holística e são compreendidos e valorizados quer pelo utente, quer por outros profissionais de saúde (Silva, 2006). Neste sentido, os métodos de avaliação das lesões músculo-esqueléticas têm sido modificados nos últimos anos. Em oposição a uma avaliação feita somente de acordo com o exame físico (incluindo testes de força muscular, mobilidade articular e avaliação dos exames complementares de diagnóstico), esta deve ainda ser acompanhada por questionários e escalas (instrumentos centrados no utente) (Puga, Lopes, & Costa, 2012). Os questionários são amplamente utilizados para coletar importantes dados clínicos, como a intensidade da dor, os níveis de qualidade de vida do paciente, a satisfação com o tratamento e a incapacidade de realizar atividades diárias (Kyte et al., 2015).

Os instrumentos de autorresposta permitem compreender a opinião do utente sobre a perceção da sua condição, bem como do tratamento na sua vida. Atualmente, o uso simultâneo de instrumentos de avaliação objetiva e instrumentos de autorresposta é comum, assegurando que são recolhidos aspetos importantes para o utente. Envolver o paciente neste processo pode ajudar a estimular um comportamento mais ativo no seu tratamento. Este tipo de instrumentos pode também ser utilizado pelos profissionais de saúde, em conjunto com o utente, para identificar o principal problema em termos de funcionalidade e de limitações nas atividades da vida diária (Kyte et al., 2015).

Existem vários instrumentos de autorresposta para a avaliação de utentes com disfunções no CAO e deteção de mudanças no seu quadro clínico ao longo do tempo, porém a maioria foi desenvolvida na língua inglesa (Bot et al., 2004; Puga et al., 2012; Roy, MacDermid, & Woodhouse, 2009). Tem havido alguns progressos através da realização das traduções e validações destas escalas e questionários para a população portuguesa. Contudo, apesar das escalas e questionários já validados para a população portuguesa e da existência de estudos que comprovam a importância das mesmas na avaliação das disfunções do CAO, não foram encontradas revisões que efetuem a análise das propriedades psicométricas dos instrumentos validados para Portugal (Bot et al., 2004; Puga et al., 2012; Roy et al., 2009; Silva, 2006). Este tipo de estudo demonstra ser essencial uma vez que as revisões sistemáticas acerca dos questionários relacionados com a saúde têm-se revelado ferramentas importantes e necessárias na seleção dos mesmos para a monitorização de utentes na prática clínica, na conceção de novos projetos de investigação, na identificação de lacunas no conhecimento da qualidade dos instrumentos de medida e como fonte de evidência sobre as propriedades psicométricas (C. Terwee et al., 2016). Além disso, permite uma visão clara e compreensiva das propriedades de medida de todos os instrumentos para inferir qual o melhor para um dado fim (C. Terwee, 2011b).

A seleção do instrumento adequado deve-se basear em grande parte na força das suas medidas psicométricas (Kyte et al., 2015). Assim levanta-se o problema: Qual/ais os instrumentos de medida de autorresposta relacionados com a funcionalidade do ombro que apresenta/m melhor/es propriedades psicométricas?

Um instrumento de medida deve revelar eficácia no que diz respeito às suas características psicométricas, uma vez que estas indicam as qualidades de utilização e eventuais fraquezas, permitindo percecionar a sua eficiência. Além disso, a seleção do instrumento utilizado deve ser realizada de forma ponderada para garantir que este tem o potencial desejado para auxiliar no raciocínio clínico, tratamento e tomada de decisão partilhada (Frost et al., 2007). Assim, na altura de selecionar um instrumento de autorresposta, o profissional de saúde deverá analisar as seguintes questões: O que pretende medir? Qual o raciocínio da avaliação? Pretende avaliar um indivíduo ou um grupo? Qual o questionário recomendado naquela situação ou qual é utilizado por outros profissionais? (Kyte et al., 2015).

Relativamente às propriedades psicométricas, existe uma falta de consenso no que diz respeito à nomenclatura, terminologia e definições. Na literatura existe uma grande quantidade de informação acerca das propriedades de medida, sendo esta por vezes divergente e inexplícita. Desta forma foi realizado um estudo Delphi que reuniu diversos *experts* para chegarem a um consenso acerca desta temática. Assim, ao longo deste projeto, será utilizada a nomenclatura proposta nesse estudo, mediante pedido de autorização de utilização do mesmo aos respetivos autores (Mokkink et al., 2010). Para além disso, atualmente o grupo "*COnsensus-based Standards for the selection of health Measurement INstruments"* (COSMIN) tem-se demonstrado relevante nas pesquisas da área da saúde para a definição das propriedades de medida e as COSMIN *checklists* têm contribuído para a revisão de escalas de avaliação relacionadas com a saúde. O grupo veio iluminar os desafios enfrentados pelos autores de escalas contemporâneas na pesquisa de evidências da qualidade dos seus instrumentos (Polit, 2015). Posto isto, segundo a terminologia proposta pela COSMIN, destacam-se as seguintes definições:

2.1 PROPRIEDADES PSICOMÉTRICAS

2.1.1 Confiabilidade

A confiabilidade compreende três propriedades: consistência interna, confiabilidade e erro de medição. Esta refere-se ao grau em que um instrumento está livre de erro e estima a medida na qual a pontuação para os utentes (que não alteraram a sua condição) é a mesma para medições repetidas sob várias condições: utilizando diferentes conjuntos de itens do mesmo instrumento de medida (consistência interna); ao longo do tempo (teste-reteste); por diferentes pessoas na mesma ocasião (inter observador); pela mesma pessoa em ocasiões diferentes (intra observador) (L. Mokkink et al., 2010; Scholtes, Terwee, & Poolman, 2011). No caso de questionários de autorresposta a confiabilidade inter observador e intra observador não é aplicável uma vez que as pontuações são definidas através da resposta do utente não sendo necessária nenhuma interpretação.

2.1.1.1 Consistência interna

É a medida na qual os itens presentes numa (sub)escala de um questionário estão correlacionados, medindo assim o mesmo constructo (L. Mokkink et al., 2010). Esta

revela-se importante para questionários que pretendam medir um único constructo utilizando vários itens. O alfa de *Cronbach* é considerado a medida adequada para a consistência interna, sendo que deverá ser calculado para cada subescala. Um baixo alfa de *Cronbach* indica falta de correlação entre os itens de uma (sub)escala, o que torna o resumo dos itens injustificada. Por outro lado um alfa de *Cronbach* muito elevado indica alto nível de correlação entre os itens da escala, ou seja, redundância num ou mais itens (C. B. Terwee et al., 2007).

2.1.1.2 Confiabilidade

A confiabilidade diz respeito ao grau em que os pacientes podem ser distinguidos uns dos outros, apesar do erro de medição (erro de medição relativo), demonstrando-se importante para fins discriminatórios caso seja necessário realizar a distinção entre os pacientes (L. Mokkink et al., 2010). Esta propriedade caracteriza-se pela proporção de variabilidade numa medida observada que é devida à variabilidade real entre indivíduos. Assim a confiabilidade poderá ser calculada através da divisão da variabilidade real entre indivíduos pela soma da variabilidade real entre indivíduos com o erro de medição. Desta forma este parâmetro apresenta-se como relativo sendo que varia sempre entre 0 e 1: quanto menor for o erro de medição maior será a confiabilidade, sendo que para 1 é considerada uma confiabilidade perfeita (Scholtes et al., 2011). O Coeficiente de Corelação Intraclasse (CCI) é o parâmetro mais adequado de confiabilidade e mais comumente utilizado, contudo poderão ser utilizados outros parâmetros como o coeficiente *Kappa* ponderado, coeficiente de *Pearson* ou coeficiente de *Spearman* (Scholtes et al., 2011; C. B. Terwee et al., 2007).

2.1.1.3 Erro de medição

Mede o erro sistemático e aleatório da pontuação de um utente que não está atribuído a verdadeiras mudanças no constructo a ser medido, podendo ser expresso pelo Erro Padrão de Medição (EPM) (L. Mokkink et al., 2010; C. B. Terwee et al., 2007). O EPM pode ser convertido na Menor Alteração Detetável (MAD). A $\text{MAD} = 1.96 \text{ x } \sqrt{2} \times EPM$, reflete a menor mudança intrapessoal na pontuação que pode ser interpretado como uma "mudança real", acima do erro de medição num indivíduo (C. B. Terwee et al., 2007). As alterações que excedam a MAD podem ser definidas como mudanças além do erro de medição. Uma outra abordagem caracteriza-se em calcular os Limites

de Concordância (LC), que correspondem à alteração média da pontuação obtida em medições repetidas +/- a MAD $\left(1.96 \text{ x } \sqrt{2} \times EPM\right)$. Os LC são utilizados muitas vezes devido à sua fácil interpretação. Para determinar a adequação do EPM, da MAD e/ou do LC estes deverão relacionar-se com a Mínima Alteração Importante (MAI) (definida na interpretabilidade). Como os erros de medição são expressos nas unidades de medidas, é impossível dar um valor para esta adequação. No entanto, é importante que o erro de medição (expresso como MAD ou LC) não seja maior do que a MAI que se quer avaliar (Schellingerhout et al., 2012).

2.1.2 Validade

A validade refere-se ao grau em que um instrumento mede o constructo que é suposto medir. Existem três tipos de validade: de conteúdo, de critério e de constructo (L. Mokkink et al., 2010; Scholtes et al., 2011).

2.1.2.1 Validade de Conteúdo

Esta propriedade examina o grau no qual o conteúdo de um questionário reflete adequadamente o constructo a ser medido (L. Mokkink et al., 2010). Assim, deverá ser especificado o objetivo da medida do questionário uma que diferentes itens podem ser válidos para objetivos diferentes: população-alvo (população para a qual foi desenvolvida o questionário), a relevância e abrangência dos itens, conceitos que o questionário pretende medir, em que medida os itens no questionário refletem áreas que são importantes para a população alvo e interpretabilidade dos itens (Scholtes et al., 2011; C. B. Terwee et al., 2007).

2.1.2.2 Validade de critério

Refere-se ao grau em que as pontuações de um instrumento particular se relacionam com o "*gold standard*" (L. Mokkink et al., 2010). De acordo com o grupo COSMIN, a validade de critério só pode ser avaliada quando existe um "*gold standard*" definido. Na perceção em saúde, nomeadamente em instrumentos de autorresposta, os investigadores do grupo COSMIN chegaram ao consenso que não existe um "*gold standard*" para este tipo de instrumentos. Enfatizam ainda que utilizar a comparação com outro instrumento como "*gold standard*" para estabelecer a validade de critério é incorreta. Se a correlação

entre os dois instrumentos comparados for baixa, não se conhece qual dos instrumentos apresenta baixa validade de critério (Scholtes et al., 2011). Embora não seja aconselhada, esta comparação é realizada por diversos autores. Assim, importa que o instrumento comparativo seja realmente considerado "*gold*".

2.1.2.3 Validade de constructo

Estima o grau no qual a pontuação do instrumento de medida é consistente com as hipóteses, baseado no pressuposto que o instrumento mede o constructo que é suposto medir (L. Mokkink et al., 2010).

As hipóteses devem ser pré-definidas e deverão abordar as relações internas esperadas, relações com a pontuação de outros instrumentos ou diferenças entre grupos relevantes, devendo ser o mais específicas possível (L. Mokkink et al., 2010; C. B. Terwee et al., 2007). No entanto não existe um consenso acerca do número de hipóteses que deverão ser testadas (Scholtes et al., 2011).

Para além do teste de hipóteses, a validade de constructo inclui ainda a validade estrutural e a validade transcultural. A validade estrutural estima o grau para o qual as pontuações do instrumento de medida refletem adequadamente a dimensão do constructo a ser medido (L. Mokkink et al., 2010). Uma análise fatorial deverá ser efetuada para confirmar o número de subescalas presentes num questionário (Schellingerhout et al., 2012). Já a validade transcultural estima o grau no qual o desempenho dos itens num instrumento traduzido e culturalmente adaptado refletem adequadamente o desempenho dos itens na versão original do instrumento (L. Mokkink et al., 2010).

2.1.3 Capacidade de Resposta

2.1.3.1 Capacidade de Resposta

Diz respeito à capacidade de um instrumento detetar mudança ao longo do tempo no constructo a ser medido (L. Mokkink et al., 2010). Por outras palavras, é a capacidade de um questionário para detetar mudanças clinicamente significativas ao longo do tempo, mesmo que essas mudanças sejam pequenas (C. B. Terwee et al., 2007).

Foi proposto um grande número de definições e métodos para avaliar a capacidade de resposta, considerando que esta se caracteriza como uma medida de eficácia longitudinal. Em analogia com validade de constructo, a correlação entre alterações das escalas de duas medidas devem estar de acordo com as hipóteses pré-definidas. Estas poderão compreender as correlações esperadas entre mudanças nas medidas ou diferenças esperadas entre "grupos conhecidos". Isto mostra a capacidade de um questionário para medir as mudanças se estas acontecerem (Schellingerhout et al., 2012; C. B. Terwee et al., 2007).

2.1.4 Interpretabilidade

Define-se como o grau em que se pode atribuir significado qualitativo para pontuações quantitativas. Por outras palavras, é o grau para o qual se podem atribuir conotações clínicas comuns e significativas à pontuação obtida na escala e respetivas mudanças (L. Mokkink et al., 2010). Para isso, os investigadores devem fornecer informações acerca de qual a alteração na pontuação que seria considerada clinicamente significativa para os utentes. A interpretabilidade pode ser expressa de certo modo pela MAI, que representa o mínimo de mudança necessária nas pontuações para esta alteração ser considerada benéfica pelo utente. A MAI deve ser definida para permitir a interpretação de alterações das escalas ao longo do tempo e os cálculos do tamanho da amostra (C. B. Terwee et al., 2007). Apesar da interpretabilidade não representar uma propriedade psicométrica, esta é ainda assim uma característica importante na avaliação de um instrumento de medição (Schellingerhout et al., 2012).

3.METODOLOGIA

3.1 OBJETIVOS

A presente revisão tem como objetivos: identificar estudos que se propõem investigar as propriedades psicométricas de instrumentos de autorresposta acerca da funcionalidade do CAO, avaliar a qualidade metodológica dos estudos, avaliar as propriedades psicométricas de cada instrumento, providenciar aos profissionais de saúde uma síntese da evidência acerca das propriedades psicométricas de cada instrumento e concluir acerca da qualidade dos instrumentos traduzidos e validados para a população portuguesa.

3.2 TIPO DE ESTUDO

Este estudo consiste numa revisão sistemática e é classificado como qualitativo, uma vez que o objetivo é recolher, analisar e compilar a informação selecionada sobre as propriedades psicométricas das escalas e questionários de autorresposta já validadas para a população portuguesa na avaliação da funcionalidade do ombro (Evans & Pearson, 2001). Para tal foi seguida a metodologia do artigo "*A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease*", contudo diferenciou-se pelo facto de não se restringir a uma só condição (Huang, Grant, Miller, Mirza, & Gagnier, 2015). Foram ainda seguidas as recomendações do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) no desenvolvimento da revisão sistemática, mais especificamente o protocolo de revisões sistemáticas de propriedades de medida desenvolvido por C. Terwee, seguindo a ideologia da COSMIN (C. Terwee, 2011b).

3.3 ESTRATÉGIA DE BUSCA

A revisão da literatura foi realizada em dois idiomas, inglês e português, pois embora a informação necessária seja relativa a Portugal esta poderia ter sido redigida na língua inglesa. Foi efetuada a pesquisa nas seguintes bases de dados: PubMed; PEDro; Google Académico; B-On e Repositório Científico de Acesso Aberto de Portugal (RCAAP).

Com as palavras-chave: "*Assessment instrument*"; "*Assessment scale*"; "*measurement scale*"; "*Outcome measure*"; "PROM"; "*Evaluation*"; "*Psychometric properties*"; "*Reliability*"; "*Validity*"; "*Responsiveness*"; "*Shoulder injury*"; "*Upper limb*"; "*Portuguese*"; "Portugal"; "Instrumento de avaliação"; "Instrumento de medida"; "Avaliação"; "Instrumentos de autorresposta"; "Propriedades psicométricas"; "Validade"; "Confiabilidade"; "Reprodutibilidade"; "Capacidade de Resposta"; "Ombro"; "Membro superior". Todos estes conceitos serão ainda conjugados através dos termos "OR" e "AND" (Bartels, 2013) (ver ANEXO I). Foram incluídos artigos/trabalhos académicos realizados até ao ano de 2015, inclusive.

3.4 CRITÉRIOS DE SELEÇÃO

Os critérios de inclusão tidos em conta foram: artigos/trabalhos académicos sobre propriedades psicométricas de questionários de autorresposta relacionados com a funcionalidade do membro superior/ombro; apresentarem pelo menos uma das características psicométricas; a escala/questionário do artigo/trabalho académico estar traduzida para português-europeu e adaptado para a população portuguesa. Por outro lado, foram excluídos: estudos em que a população fosse predominantemente constituída por crianças (<18 anos) e artigos/trabalhos académicos sem acesso ao documento completo.

3.5 INSTRUMENTOS

As revisões sistemáticas das propriedades de medida são úteis e fornecem evidência para a seleção do instrumento de medida com mais qualidade para um objetivo específico (L. B. Mokkink et al., 2009). Numa revisão sistemática não devem ser tidos em conta apenas os resultados dos estudos incluídos mas também a qualidade metodológica da própria revisão. A avaliação da qualidade metodológica de um estudo e a avaliação da qualidade dos instrumentos incluídos são dois aspetos distintos e deverão ser realizados separadamente nas revisões sistemáticas (C. B. Terwee et al., 2012).

3.5.1 Avaliação da Qualidade Metodológica

Se a qualidade metodológica de um estudo sobre as propriedades de medida de um instrumento específico for apropriada, os resultados poderão ser utilizados para a

avaliação da qualidade do instrumento. Contudo, quando a qualidade metodológica de um estudo é inadequada, os resultados não são confiáveis e a qualidade do instrumento em estudo não é clara. Alguns autores de revisões sistemáticas sobre propriedades de medida avaliaram a qualidade metodológica dos estudos incluídos. Contudo foram utilizados diferentes métodos para avaliarem a qualidade metodológica (C. B. Terwee et al., 2012). Para atenuar estas diferenças foi desenvolvida a COSMIN *checklist* para a avaliação da qualidade metodológica dos estudos sobre propriedades de medida (L. B. Mokkink et al., 2010). Posteriormente foi desenvolvido um sistema de pontuação para calcular a qualidade metodológica por cada propriedade de medida presente na COSMIN *checklist*. Neste ponto foi desenvolvida a COSMIN *checklist* modificada (com escala de 4 pontos) (C. B. Terwee et al., 2012; C. Terwee, 2011a). Nesta versão as quatro opções de resposta para cada item da *checklist* foram definidas como "excelente", "bom", "médio" e "fraco". A pontuação final da qualidade metodológica do estudo por cada propriedade de medida é obtida utilizando a avaliação mais baixa obtida em qualquer item incluído nessa propriedade ("a pior pontuação conta"). Por exemplo, se um dos itens contidos na propriedade de confiabilidade é pontuada como "fraca", a avaliação da qualidade metodológica no estudo é considerada fraca. O objetivo deste sistema é obter uma pontuação global da qualidade metodológica por propriedade de medida para um determinado estudo. Desta forma as pontuações da qualidade metodológica para diferentes estudos não deverão ser combinadas. Por exemplo, se a revisão sistemática incluir três estudos sobre a confiabilidade do mesmo instrumento de medida, a qualidade metodológica de cada estudo deverá ser avaliada separadamente (C. B. Terwee et al., 2012).

Assim, tal como recomendado por Terwee, nesta revisão sistemática foi utilizada a COSMIN *checklist* modificada como instrumento de avaliação da qualidade metodológica dos estudos incluídos (C. Terwee, 2011b). Esta avaliação foi realizada por duas revisoras, na qual foram elaboradas, de forma independente, uma COSMIN *checklist* modificada para cada um dos artigos selecionados. Em caso de discordância entre as duas revisoras, a resolução do problema foi alcançada através de consenso. Quando este não era possível, uma terceira pessoa foi consultada para resolver a discordância (C. Terwee, 2011b).

3.5.2 Avaliação das Propriedades Psicométricas

Como forma de avaliar as propriedades psicométricas da literatura selecionada, foi utilizada uma escala de classificação proposta inicialmente por Terwee et al. (C. B. Terwee et al., 2007). Contudo foi aplicada uma classificação modificada por Schellingerhout et al., tendo sido esta utilizada igualmente no estudo de Huang et al. (Huang et al., 2015; Schellingerhout et al., 2012). Esta classificação modificada apresenta como vantagem, relativamente à proposta inicial, a utilização da nomenclatura definida pela COSMIN (nomenclatura adotada para a presente revisão) (L. Mokkink et al., 2010). Esta escala inclui critérios de qualidade para: a consistência interna, a confiabilidade, o erro de medição, a validade de conteúdo, a validade de critério, a validade de constructo (validade estrutural e teste de hipóteses) e capacidade de resposta. Cada critério é classificado através de sinais como o positivo (+), o indeterminado (?), o negativo (-) e o (na) quando não existe informação da literatura (Schellingerhout et al., 2012). Estes critérios encontram-se descritos na Tabela 1.

3.5.3 Métodos de síntese

Para sintetizar todos os dados recolhidos e chegar a uma categorização geral da evidência, foi utilizado um método proposto por Terwee (C. Terwee, 2011b). Neste método a síntese dos diferentes estudos é realizada através da combinação dos seus resultados, recolhendo o número e a qualidade metodológica dos estudos (através dos critérios COSMIN) com a consistência da classificação da evidência psicométrica (avaliação das propriedades psicométricas) baseado nos níveis de evidência propostos pelo *Cochrane Back Review Group* (van Tulder, Furlan, Bombardier, & Bouter, 2003). Assim as propriedades psicométricas são classificadas como positiva (+), indeterminada (?) ou negativa (-) que, aquando da combinação dos estudos corresponde aos seguintes níveis de evidência: forte, moderado, limitado, conflituoso e desconhecido (Tabela 2) (C. Terwee, 2011b).

Tabela 1- Critérios de qualidade para avaliação das propriedades psicométricas

Propriedade	Classificação	Critério de qualidade
Confiabilidade		
Consistência Interna	+	Alfa(s) de *Cronbach* ≥0.70
	?	Dimensão desconhecida OU alfa (s) de *Cronbach* não determinado(s)
	-	Alfa(s) de *Cronbach* <0.70
Erro de Medição	+	MAI> MAD OU MAI fora do LC
	?	MAI não definido
	-	MAI≤MAD OU MAI igual ou dentro do LC
Confiabilidade	+	CCI/ *kappa* ponderado ≥0.70 OU *Pearson's* r≥0.80
	?	CCI/ *kappa* ponderado/ *Pearson's* não determinado
	-	CCI/ *kappa* ponderado <0.70 OU *Pearson's* r<0.80
Validade		
Validade de conteúdo	+	A população-alvo considera todos os itens do questionário como relevante E considera o questionário completo
	?	Sem população alvo envolvida
	-	A população alvo considera itens no questionário irrelevantes E considera o questionário incompleto
Validade de constructo- Validade estrutural	+	Fatores devem explicar, pelo menos, 50% da variância
	?	Explicação da variância não mencionada
	-	Fatores explicam <50% da variância
Validade de constructo- Teste de hipóteses	+	(Correlação com um instrumento de medição com o mesmo constructo ≥ 0.50, OU, pelo menos, 75% dos resultados estão de acordo com as hipóteses) E correlação com construções relacionadas é maior do que com construções independentes.
	?	Apenas correlações determinadas com construções independentes
	-	Correlação com um instrumento de medição com o mesmo constructo <0.50 OU <75% dos resultados estão de acordo com as hipóteses OU correlação com as construções relacionadas é mais baixa do que com as construções independentes
Validade de critério	+	Argumentos convincentes que o "*gold standard*" é realmente "*gold*" E correlação com o "*gold standard*" ≥0.7
	?	Sem argumentos que o "*gold standard*" é realmente "*gold"* OU design ou método duvidoso
	-	Correlação com o padrão <0.70, apesar de método e *design* adequado
Capacidade de resposta		
Capacidade de resposta	+	(Correlação com um instrumento de medição com o mesmo constructo ≥ 0.50 OU pelo menos, 75% dos resultados estão de acordo com as hipóteses OU AAC≥0.70) E (correlação com construções relacionadas é maior do que com construções independentes)
	?	Apenas correlações determinadas com construções independentes
	-	Correlação com um instrumento de medição com o mesmo constructo <0.50 OU < 75% dos resultados estão de acordo com as hipóteses OU AAC<0.70 OU correlação com construções relacionadas é menor do que com construções independentes.

Legenda: CCI- Coeficiente de Correlação Intraclasse; MAI- Mínima alteração importante; MAD-Mínima Alteração Detetável; LC-Limites da concordância; AAC- Área abaixo da Curva ROC (*Receiver Operating Characteristic*)

Tabela 2- Níveis de evidência para a qualidade geral das propriedades de medida

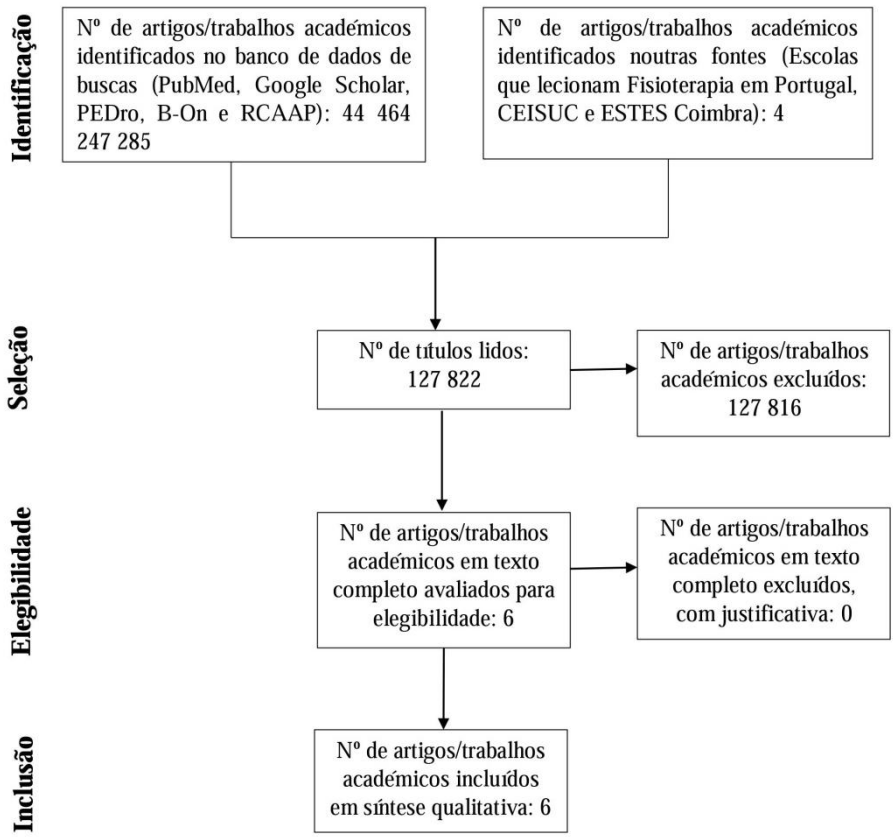
Nível de Evidência	Classificação*	Critério
Forte	+++ ou ---	Resultados consistentes em múltiplos estudos com boa qualidade metodológica OU num estudo com qualidade metodológica excelente
Moderado	++ ou --	Resultados consistentes em múltiplos estudos com qualidade metodológica média OU num estudo com boa qualidade metodológica
Limitado	+ ou -	Um estudo com qualidade metodológica média
Conflituoso	+/-	Resultados conflituosos
Desconhecido	?	Apenas estudos com fraca qualidade metodológica

*O sinal mais (+) indica um resultado positivo; o sinal menos (-) indica um resultado negativo

4.RESULTADOS

Após a identificação de estudos, foram lidos cerca de 127822 títulos, sendo que foram incluídos apenas 6 estudos disponíveis em texto completo que cumpriam os critérios de elegibilidade (Imagem 1). Assim, nesta revisão foram incluídos 6 estudos sobre as propriedades psicométricas de 6 instrumentos de autorresposta de avaliação da funcionalidade do ombro. Os instrumentos analisados foram: *Disabilities of the Arm Shoulder and Hand* (DASH); *Neck and Upper Limb Index* (NULI-20); *Shoulder Pain And Disability Index* (SPADI); *Shoulder Rating Questionnaire* (SRQ-PT); *Upper Extremity Functional Index* (UEFI); *Weelchair User's Shoulder Pain Index* (WUSPI) (Clara, 2001; Duarte, 2002; Guerreiro, Proença, Moura, & Cartucho, 2011; Matias, 2010; Melo, 2002; Santos & Gonçalves, 2006).

Imagem 1- Fluxograma da pesquisa da literatura



A caracterização dos estudos incluídos, bem como a descrição dos questionários, encontra-se resumida na Tabela 3 e na Tabela 4, respetivamente. Posto isto, para cada estudo incluído nesta revisão foram realizadas 2 avaliações, conduzindo a 12 avaliações independentes da metodologia que, após consenso, originaram os resultados apresentados na Tabela 5 (Ver ANEXO II). A avaliação de cada propriedade psicométrica segundo os critérios de qualidade descritos na Tabela 1, encontram-se igualmente contemplados na Tabela 5. Os resultados obtidos foram sintetizados segundo os níveis de evidência (Tabela 2) e resumidos na Tabela 6.

Tabela 3- Caracterização dos estudos

Estudo	Autor	Ano	Amostra	Idades	Caracterização da amostra
Adaptação e validação cultural da versão portuguesa do *Disabilities of the Arm Shoulder and Hand*-DASH	Santos, J. Gonçalves, R.	2006	Amostra de conveniência de 54 adultos; (2 subamostras de 34 e 54 indivíduos)	Indivíduos adultos (maiores de 18 anos)	Indivíduos com disfunções no membro superior
Tradução e Adaptação cultural do *Neck and Upper Limb Index* para a língua portuguesa	Matias, S.	2010	Amostra de 81 indivíduos; (3 subamostras de 81, 42 e 41 indivíduos)	48.46 ± 9.64 anos	Cervical n=4 (6.2%); Cervical e ombro(s) n=16 (19.8%); Ombro(s) n=23 (28.4%); Cotovelo(s) n=23 (28.4%); Punho(s) e Mão(s) n=12 (14.8%); Artrose/artrite n=5 (6.2%); Tendinopatia n=52 (64.2%); Bursite n=2 (2.5%); Nevralgia n=2 (2.5%); Algia de origem não especificada n=17 (21.0%)
Validação intercultural do *Shoulder Pain and Disability Index*- SPADI	Duarte, A.	2002	Amostra de conveniência de 29 indivíduos; (2 subamostras de 10 e 29 indivíduos)	51.03 ± 13.073 anos	Osteoartrite n=2 (6.9%); Tendinopatia do supraespinhoso n=9 (31%); Rutura do tendão supraespinhoso n=1 (3.4%); Tendinopatia coifa dos rotadores n=1 (3.4%); Rutura coifa dos rotadores n=2 (6.9%); Capsulite adesiva n=4 (13.8%); Fraturas n=5 (17.2%); Outro n=5 (17.2%)
Adaptação Transcultural do *Shoulder Rating*	Guerreiro, J.; Proença,	2011	Amostra consecutiva de 55 indivíduos	Indivíduos adultos (maiores	Conflito subacromial n=28 (50.9%); Rutura da coifa dos rotadores n=5 (9.1%); Luxação recidivante do ombro n=5

Questionnaire para a língua portuguesa (SRQ-PT): Tradução; Validação; Análise da Consistência Interna e Replicabilidade	I.; Moura, N.; Cartucho, N.			de 18 anos)	(9.1%); Luxação da articulação acrómio-clavicular n=4 (7.3%); Capsulite adesiva n=3 (5.5%); Rutura parcial do supraespinhoso n=2 (3.6%); Artrose da articulação acrómio-clavicular n=2 (3.6%); Tendinopatia calcificada n=2 (3.6%); Condição pós-cirúrgica n=1 (1.8%); Artrose da articulação esterno-clavicular n=1 (1.8%); Sem diagnóstico definido n=2 (3.3%)
Upper Extremity Functional Index – Adaptação cultural e linguística	Melo, F.	2002	Amostra de conveniência de 28 indivíduos (2 subamostras de 28 e 12 indivíduos)	39.82 ± 13.48 anos	Localização da lesão: Ombro n=15 (53.6%); Cotovelo n=8 (28.6%); Punho/mão n=5 (17.9%)
Validação intercultural do *Wheelchair User's Shoulder Pain Index*	Clara; F.	2001	Amostra de 28 indivíduos (3 subamostras de 28, 15 e 12 indivíduos)	39.61 ± 11.06 anos	Dor no ombro antes da utilização da cadeira de rodas n=2 (7.1%); Dor no ombro desde que usa a cadeira de rodas n=19 (67.9%); Neste momento n=16 (57.1%);

Tabela 4- Descrição dos instrumentos

Instrumento	Referência	Ano	Segmento corporal avaliado	Nº de itens e dimensões	Dimensões	Opções de resposta	Sistema de pontuação
DASH	(Santos & Gonçalves, 2006)	Estudo publicado, ano 2006	Membro superior	30 itens, 2 dimensões (c/2 módulos opcionais c/4 itens cada)	Sintomas e incapacidade; Módulos opcionais (Desporto/Música e Trabalho)	1 a 5	Orientação negativa de 0 (máxima funcionalidade) a 100 (máxima incapacidade)
NULI-20	(Matias, 2010)	Estudo não publicado, ano 2010	Membro superior e Cervical	20 itens, 5 dimensões	Trabalho, Atividades físicas, Sono, Impacto psicossocial, Efeitos iatrogénicos	1 a 7 (0 se não aplicável)	Orientação negativa de 1 (ausência de incapacidade) a 7 (máxima incapacidade)

SPADI	(Duarte, 2002)	Estudo não publicado, ano 2002	Ombro	13 itens, 2 dimensões	Dor e Incapacidade	EVA 10 cm para cada item	Orientação negativa de 0 (ausência de dor/ ausência de dificuldade) a 10 (pior dor possível/máxima dificuldade)
SRQ-PT	(Guerreiro et al., 2011)	Estudo publicado, ano 2011	Ombro	21 itens, 6 dimensões	Avaliação global, Dor, Atividades da vida diária, Atividades desportivas e de lazer, Satisfação global, Trabalho	EVA 10 cm; pontuação de 1 a 5; na última questão o utente classifica por ordem de preferência duas áreas onde gostaria de ver melhorado o seu desempenho	Orientação positiva de 0 (muito mal) a 10 (muito bem); De 1 (pior) a 5 (melhor). Total entre 17 e 100 pontos
UEFI	(Melo, 2002)	Estudo não publicado, ano 2002	Membro superior	20 itens, 1 dimensão	Unidimensional	Pontuação de 0 a 4	Orientação positiva de 0 (máxima incapacidade) a 80 (máxima funcionalidade)
WUSPI	(Clara, 2001)	Estudo não publicado, ano 2001	Ombro	15 itens, 1 dimensão	Unidimensional	EVA 10 cm para cada item	Orientação negativa de 0 (ausência de dor) a 150 (pior dor alguma vez sentida)

Disabilities of the Arm Shoulder and Hand (DASH)

Relativamente ao questionário do DASH foi encontrado apenas um estudo relativo à avaliação das propriedades psicométricas (Santos & Gonçalves, 2006). Neste foi avaliado a consistência interna, a confiabilidade, o teste de hipóteses, a validade transcultural e a validade de critério. A consistência interna apresenta um alfa de *Cronbach* de 0.95, classificando-se com sinal positivo, contudo, segundo a classificação COSMIN, esta apresenta uma qualidade metodológica fraca. Desta forma, para esta propriedade a evidência é desconhecida. Relativamente à confiabilidade esta foi calculada através da correlação de *Pearson*, com um valor de 0.886, sendo cotada com sinal positivo. No entanto, esta apresenta uma qualidade metodológica média. Desta forma o nível de evidência para esta propriedade é limitado positivo. No teste de hipóteses os autores analisaram as correlações existentes entre o DASH, a severidade da dor e o grau de incapacidade, corroborando o que inicialmente tinham conjeturado, obtendo valores de correlação de 0.49 para a severidade da dor e de 0.54 para o grau de incapacidade. Segundo a classificação COSMIN, o teste de hipóteses apresenta uma qualidade metodológica fraca, adquirindo assim um nível de evidência desconhecido. Relativamente à validade transcultural apresenta uma qualidade metodológica fraca dado que foram realizados todos os passos de tradução descritos na literatura, resultando a versão portuguesa pré-final. Esta versão foi submetida ao Comité da *American Academy of Orthopaedic Surgeons* para apreciação. Por fim foi realizado um pré-teste, no qual um painel de indivíduos procedeu ao preenchimento do questionário DASH e à compreensão dos itens. Para a validade de critério os autores correlacionaram as pontuações obtidas no DASH com as pontuações obtidas no *Short-form Health Survey 36-Item* (SF-36), obtendo uma correlação negativa e significativa, não apresentando o valor desta correlação. Contudo a qualidade metodológica para a validade de critério é classificada como boa.

Neck and Upper Limb Index (NULI-20)

Relativamente ao questionário NULI-20 foi encontrado apenas um estudo alusivo à avaliação das propriedades psicométricas, no qual foi avaliado a consistência interna, a confiabilidade, a validade de conteúdo, o teste de hipóteses, a validade transcultural, a validade de critério e a capacidade de resposta (Matias, 2010). Pela análise dos

resultados, verificou-se que a consistência interna da versão portuguesa tem um valor de alfa de *Cronbach* igual a 0.92 para o questionário global, um valor de classificação positivo. Contudo, segundo a COSMIN, a consistência interna deste estudo é considerada de fraca qualidade metodológica, tendo portanto um nível de evidência desconhecido. No estudo da confiabilidade foi calculado o CCI, que obteve o valor de 0.83 tendo sido cotada com sinal positivo. Porém, na classificação COSMIN, esta é considerada de média qualidade metodológica, respeitante a um nível de evidência limitado positivo. Quanto à validade de conteúdo, os indivíduos entrevistados no estudo consideraram a versão portuguesa do NULI-20 clara, compreensível e adequada à sua situação clínica, classificando o questionário como breve, de fácil e rápida resposta, compreensível e útil. Foi igualmente unânime a opinião de que a linguagem utilizada era simples, clara e coloquial, obtendo assim uma qualidade metodológica classificada como excelente e um nível de evidência forte positivo. O teste de hipóteses foi avaliado através do coeficiente de correlação de *Spearman*, segundo o qual a pontuação global do questionário foi de 0.612, correspondendo a um sinal positivo. Todavia, segundo a COSMIN, a qualidade metodológica é média e o nível de evidência limitado positivo. Relativamente à validade transcultural foi realizada a equivalência semântica obtida pela tradução, retroversão, obtenção de uma versão de consenso e análise da qualidade da tradução realizada por dois clínicos. Apesar destes procedimentos, esta foi avaliada com qualidade metodológica fraca segundo a COSMIN. A validade de critério foi avaliada através do coeficiente de correlação de *Spearman* e este varia entre os valores de -0.340 e -0.688, sendo classificada com sinal negativo. Por outro lado, a qualidade metodológica é considerada boa e com nível de evidência moderado negativo. A capacidade de resposta foi avaliada através da medida estatística *standardized effect size* (ES) apresentando-se com ES =0.95, ou seja, um sinal positivo. Relativamente à qualidade metodológica, no NULI-20 a capacidade de resposta é fraca, correspondendo a um nível de evidência desconhecido.

Shoulder Pain And Disability Index (SPADI)

Quanto ao questionário SPADI foi encontrado apenas um estudo relativo à avaliação das suas propriedades (Duarte, 2002). Neste foram contempladas propriedades psicométricas como a coerência interna, confiabilidade, validade de conteúdo, teste de hipóteses, validade de critério e a capacidade de resposta. Segundo a análise dos

resultados temos que a consistência interna da versão portuguesa, calculada pelo alfa de *Cronbach*, apresenta um valor de 0.75 para a dimensão dor e de 0.84 para a dimensão atividade funcional, sendo ambas cotadas com sinal positivo. Contudo, apresenta uma qualidade metodológica fraca de acordo com a classificação COSMIN, resultando num nível de evidência desconhecido. Quanto à confiabilidade foi aplicada a correlação de *Pearson*, sendo obtidos valores de 0.898 para a dimensão da dor e de 0.861 para a dimensão atividade funcional, ambas positivas. No entanto, segundo a classificação da COSMIN, a qualidade metodológica é fraca, resultando assim num nível de evidência desconhecido. Quanto à validade de conteúdo, os indivíduos entrevistados foram questionados apenas acerca da compreensão da primeira pergunta de cada dimensão, sendo assim classificada com metodologia fraca e com nível de evidência desconhecido. O teste de hipóteses foi calculado pela análise das correlações das diferentes dimensões e a idade, a severidade da dor e o tempo de doença. Relativamente às correlações entre a pontuação do SPADI e a idade verificaram-se valores de 0.392 para a dimensão dor e de 0.282 para a dimensão atividade funcional; para a severidade da dor verificaram-se valores de 0.730 para a dimensão dor e de 0.490 para a dimensão atividade funcional; para o tempo de doença verificaram-se valores de 0.494 para a dimensão dor e de 0.532 para a dimensão atividade funcional. Os valores que se apresentam abaixo de 0.50 foram classificados com sinal negativo, contrariamente aos valores superiores, cotados com sinal positivo. Esta propriedade tem um nível de qualidade metodológica fraca segundo a COSMIN, sendo o grau de evidência desconhecido. Quanto à validade de critério, esta foi avaliada através da correlação das pontuações obtidas no SPADI com as do SF-36. Dentro da dimensão dor no SPADI foram obtidos valores de correlação que variam de -0.154 (função social) a -0.655 (dor física); dentro da dimensão da atividade funcional estes variam entre –0.087 (função social) e -0.801 (dor física). Todas as dimensões foram cotadas com sinal negativo à exceção da dimensão "dor física", tendo sido classificada com sinal positivo. Quanto a esta propriedade, segundo a classificação COSMIN, apresenta qualidade metodológica fraca, tendo um nível de evidência desconhecido. Relativamente à capacidade de resposta, esta foi dada pela correlação dos valores médios das diferenças verificadas nos dois momentos de avaliação entre a pontuação do SPADI, do SF-36 e as amplitudes de movimento, não se tendo verificado qualquer tipo de correlações. A qualidade metodológica foi classificada como fraca, sendo o nível de evidência para esta propriedade desconhecido.

Shoulder Rating Questionnaire (SRQ-PT)

Referente ao questionário do SRQ-PT foi encontrado apenas um estudo alusivo à avaliação das propriedades psicométricas, no qual foi avaliada a consistência interna, a confiabilidade, a validade de conteúdo e a validade transcultural (Guerreiro et al., 2011). Para a consistência interna foi calculado o alfa de *Cronbach*, que obteve um valor de 0.91 para o questionário total, sendo classificada com sinal positivo. Contudo, segundo a COSMIN a qualidade metodológica é fraca, apresentando um nível de evidência desconhecido. Quanto à confiabilidade foi utilizado o coeficiente de correlação de *Spearman*, obtendo para o questionário total um valor de 0.90, ou seja, um sinal positivo. Todavia, de acordo com a classificação COSMIN a qualidade metodológica é de fraca e o nível de evidência desconhecido. A equivalência semântica e de conteúdo entre as versões de língua inglesa e de língua portuguesa, foram efetuadas através duma tradução e retroversão cega, discussões informais acerca das imagens conceptuais de funcionalidade e a sua aplicabilidade à cultura portuguesa. Este processo foi levado a cabo por um *focus group* composto por profissionais de saúde. O processo de validação de conteúdo levou ainda em conta a opinião dos respondentes, sendo classificada como excelente na qualidade metodológica e com um nível de evidência forte positivo. Contrariamente, a validade transcultural é de fraca qualidade metodológica.

Upper Extremity Functional Index (UEFI)

Relativo ao questionário UEFI foi encontrado um estudo relativo à avaliação das propriedades psicométricas, no qual foi avaliada a consistência interna, a confiabilidade, a validade de conteúdo e a validade transcultural (Melo, 2002). Os resultados dos testes de consistência interna demonstram que a versão portuguesa do UEFI apresenta um valor de alfa de *Cronbach* de 0.93, apresentando um sinal positivo. Contudo, segundo a COSMIN esta propriedade é classificada como fraca na sua qualidade metodológica, que consequentemente terá um nível de evidência desconhecido. A confiabilidade foi analisada individualmente para cada um dos itens. Os valores para a correlação de *Pearson* variam entre 0.61 (Item a – "Fazer qualquer trabalho habitual, trabalho em casa ou atividades escolares") e 0.972 (item b – "Realizar os seus passatempos habituais, atividades recreativas ou desportivas"). Referente à classificação da COSMIN esta é considerada de fraca qualidade metodológica e com um nível de evidência

desconhecido. Relativamente à validade transcultural foi efetuado o processo de equivalência semântica obtida pela tradução, retroversão por duas tradutoras bilingues, obtenção da versão de consenso, aplicação da versão pré-final e de um teste de compreensão. Esta propriedade apresenta uma qualidade metodológica fraca segundo a COSMIN. Quanto à validade de conteúdo resultou o consenso de que o questionário é breve, de fácil e rápida resposta, compreensível, útil e adequado à população a que se dirige. Acresce que, foi igualmente unânime a opinião de que a linguagem utilizada é simples, clara e coloquial. Desta forma esta propriedade foi classificada com excelente qualidade metodológica e de nível de evidência forte positivo.

Weelchair User's Shoulder Pain Index (WUSPI)

Relativamente ao questionário do WUSPI foi encontrado apenas um estudo alusivo à avaliação das propriedades psicométricas (Clara, 2001). Neste foi avaliado a consistência interna, a confiabilidade, a validade de conteúdo, o teste de hipóteses, a validade transcultural, a validade de critério e a capacidade de resposta. Da análise dos resultados verificou-se que a consistência interna da versão portuguesa apresenta um valor de alfa de *Cronbach* de 0.907 (sinal positivo). No entanto, segundo a classificação COSMIN, é considerada como fraca na sua qualidade metodológica. Segundo estes dados, considera-se que o nível de evidência desta propriedade é desconhecido. Quando analisada a sua confiabilidade, o WUSPI apresenta um valor de correlação de *Pearson* de 0.998, obtendo, assim, um sinal positivo. Por outro lado, o nível de qualidade metodológica é fraco. O grau de evidência desta propriedade psicométrica considera-se como desconhecido. Relativamente à validade de conteúdo, foi unânime que este instrumento é de fácil resposta, rápido, de simples compreensão e útil, justificando-se assim a sua existência para aqueles aos quais se dirige. Este instrumento de medida apresenta qualidade metodológica excelente e um nível de evidência forte positivo. No teste de hipóteses, os autores analisaram as correlações existentes entre o WUSPI e os dados sociodemográficos, os dados relativos à utilização de cadeira de rodas e os dados relativos à sintomatologia dolorosa no ombro sendo que apenas se verificaram correlações entre o WUSPI e alguns aspetos dos dados relativos à sintomatologia dolorosa do ombro. Assim, para o teste de hipóteses, a qualidade metodológica segundo a COSMIN é fraca e o nível de evidência caracteriza-se como desconhecido. A validade transcultural foi realizada após a retroversão do referido instrumento para a língua

original, assegurada por tradutor bilingue da língua portuguesa para a língua inglesa, e consequente comparação com a versão original através de um painel de indivíduos. A qualidade metodológica deste estudo considera-se como fraca e o nível de evidência como desconhecido. Em relação à validade de critério, os autores, através da correlação de *Pearson*, correlacionaram o WUSPI com a escala de severidade da dor no ombro, as amplitudes de movimento ativo do ombro e o estado de saúde utilizando o SF-36. Relativamente à severidade da dor esta apresenta uma correlação com o WUSPI de 0.929 para o ombro direito e de 0.825 para o ombro esquerdo. No que respeita às amplitudes de movimento ativo do ombro estas apresentam correlações negativas com o WUSPI, tendo apresentado valores entre -0.416 (extensão) e -0.857 (rotação medial) para o ombro direito e valores entre -0.416 (extensão) e -0.912 (rotação medial) para o ombro esquerdo. Todas as dimensões foram cotadas com sinal positivo com exceção das componentes de extensão de ambos os ombros. Em relação ao SF-36 não foram verificadas correlações significativas com o WUSPI. A qualidade metodológica desta propriedade psicométrica caracteriza-se como fraca e o nível de evidência como desconhecido. No que concerne à capacidade de resposta, esta foi avaliada através da análise das pontuações obtidas em três momentos de aplicação: início do estudo (t_0), duas semanas após (t_1) e quatro semanas após o início (t_2). Os autores analisaram as correlações entre as médias das alterações encontradas no WUSPI com as médias das alterações encontradas na severidade da dor, nas amplitudes de movimento ativo do ombro e no estado de saúde entre os momentos t_0-t_1 e t_1-t_2. Assim as correlações existentes para a severidade da dor foram sempre mais elevadas no que respeita ao ombro esquerdo apresentando valores de 0.474 entre t_0-t_1 e de 0.511 entre t_1-t_2. As correlações existentes para as amplitudes de movimento ativo do ombro foram igualmente positivas variando entre 0.380 e 0.759 para t_0-t_1 e entre 0.165 e 0.733 para t_1-t_2. Para as médias das alterações verificadas no estado de saúde avaliado através do SF-36 não se verificaram correlações significativas com as médias das alterações do WUSPI. Para valores iguais ou superiores a 0.50, foi atribuído sinal positivo. Já para valores inferiores foi utilizado o sinal negativo. A qualidade metodológica deste estudo caracteriza-se como fraca e o nível de evidência desconhecido.

Tabela 5- Resultados

Escala	Consistência interna		Confiabilidade		Erro de mediç ão	Validade de conteúdo		Validade estrutural	Teste de hipóteses		Validade de critério		Capacidade de resposta	
	P	M	P	M		P	M		P	M	P	M	P	M
DASH	+	Fraco	+	Médio	na	na	na	na	+	Fraco	na	Bom	na	na
NULI-20	+	Fraco	+	Médio	na	+	Excelente	na	+	Médio	-	Bom	+	Fraco
SPADI	+	Fraco	+	Fraco	na	+	Fraco	na	+	Fraco	-	Fraco	-	Fraco
SRQ-PT	+	Fraco	?	Fraco	na	+	Excelente	na	na	na	na	na	na	na
UEFI	+	Fraco	+	Fraco	na	+	Excelente	na	na	na	na	na	na	na
WUSPI	+	Fraco	+	Fraco	na	+	Excelente	na	-	Fraco	-	Fraco	+	Fraco

Legenda: P- avaliação da propriedade psicométrica segundo os critérios descritos na Tabela 1; M- avaliação da metodologia do estudo segundo a *checklist* da COSMIN; (na)- não aplicável

Tabela 6- Níveis de evidência

Escala	Consistência interna	Erro de medição	Confiabilidade	Validade de conteúdo	Validade estrutural	Teste de hipóteses	Validade de critério	Capacidade de resposta
DASH	?	na	+	na	na	?	na	na
NULI-20	?	na	+	+++	na	+	--	?
SPADI	?	na	?	?	na	?	?	?
SRQ-PT	?	na	?	+++	na	na	na	na
UEFI	?	na	?	+++	na	na	na	na
WUSPI	?	na	?	+++	na	?	?	?

Legenda: (+)- nível de evidência limitado; (+++)- nível de evidência forte positivo; (--)- nível de evidência moderado negativo; (?)- nível de evidência desconhecido; (na)- não aplicável

5.DISCUSSÃO

O DASH tem sido utilizado na avaliação de qualquer condição do membro superior, no entanto este é frequentemente utilizado em condições relacionadas com o ombro. Num estudo no qual realizaram a comparação de dois tipos de intervenção (estiramento progressivo estático mais terapia tradicional versus terapia tradicional) no tratamento da capsulite adesiva do ombro, a forma de avaliação dos resultados foi com recurso ao questionário DASH (Ibrahim, Donatelli, Hellman, & Echternach, 2013). Noutro estudo realizado em indivíduos com dor no ombro não traumática, o DASH foi utilizado para verificar a efetividade de um tratamento para pontos de gatilho no alívio da dor (Bron et al., 2011). Numa revisão sistemática, alguns artigos incluídos argumentam que o DASH deve ser descartado dos processos de avaliação do ombro uma vez que não é específico para essa articulação e que, geralmente, os instrumentos específicos são mais sensíveis que os gerais. No entanto, os autores defendem que as propriedades psicométricas do DASH são tão boas ou melhores do que as das escalas específicas do ombro (Roy et al., 2009).

Relativamente aos resultados do nosso estudo, a avaliação das propriedades psicométricas do DASH apresentou uma pontuação positiva para todas as componentes segundo os critérios estabelecidos na Tabela 1.

Contudo, no que diz respeito à metodologia do estudo referente ao DASH (ANEXO II), na avaliação da consistência interna, esta caracterizou-se como fraca, uma vez que não foi calculado para cada dimensão o valor de alfa de *Cronbach*. Para além disso, a análise fatorial não foi realizada e não referem outro estudo, bem como a ausência do cálculo do ajuste estatístico a nível global. Já a confiabilidade classificou-se como média, considerando que não foram descritas as condições de teste nem o estado dos pacientes ao longo do tempo de aplicação do mesmo. Uma outra componente a considerar nesta propriedade foi o reduzido tamanho da amostra. Em relação à validade de conteúdo, esta não foi avaliada no nosso estudo uma vez que os dados apresentados no estudo eram subjetivos. Nesta propriedade os autores apenas referiram que, para além do painel de peritos, não se verificou qualquer efeito chão/teto no DASH

mostrando, assim, a validade de conteúdo da medida. Quanto ao teste de hipóteses, este foi considerado fraco dado que não foi fornecida uma descrição adequada do instrumento comparador. No que concerne à validade de critério, esta foi classificada como boa, sendo que não foi fornecida evidência que o critério utilizado poderia ser considerado um *gold standard*. Além disso, esta propriedade psicométrica não foi avaliada segundo os critérios de qualidade uma vez que os autores não apresentaram o valor da correlação entre o DASH e o SF-36, referindo apenas que os valores se correlacionavam negativa e significativamente na quase totalidade dos itens. De forma a colmatar estas falhas metodológicas, foram solicitados dados complementares ao autor correspondente do artigo, no qual não foi obtida qualquer resposta.

O NULI-20 é um instrumento que avalia o estado funcional de trabalhadores com lesões músculo-esqueléticas (ao nível da cervical e do membro superior). Este questionário permite avaliar a cervical e o membro superior ou apenas uma das estruturas dado que apresenta a opção de resposta "não aplicável". Ainda que não seja recorrente a sua utilização, no estudo de validação deste instrumento para a população portuguesa, cerca de 28.4% da amostra incluída apresentava alterações ao nível do ombro (Matias, 2010). Para além disso, foi realizado um estudo que se propôs avaliar a efetividade de um protocolo de exercícios de relaxamento muscular em músicos com alterações músculo-esqueléticas ao nível do ombro. Desta forma, o questionário NULI-20 foi aplicado com vista à perceção da redução da dor e desconforto nos músculos do CAO, nomeadamente ao nível dos trapézios e deltóides (Rodrigues, Santana, & Pinheira, 2014).

Relativamente aos resultados da nossa revisão, aquando da avaliação das propriedades psicométricas, estas foram todas avaliadas com classificação positiva exceto a validade de critério que foi pontuada como negativa uma vez que a correlação com o *gold standard* (*Short-form Health Survey 12-Item* (SF-12)) apresentava um valor inferior a 0.7.

Na avaliação da metodologia do estudo referente ao NULI-20 (ANEXO II), a consistência interna evidenciou-se como fraca uma vez que a análise fatorial não foi realizada e não referem outro estudo, bem como a ausência do cálculo do ajuste estatístico a nível global. Em relação à confiabilidade, esta foi classificada como média devido ao tamanho da amostra reduzido. Já a validade de conteúdo foi considerada

excelente. No que toca ao teste de hipóteses, foi tido como um nível médio, dado que apenas foram dadas algumas informações sobre as propriedades de medida do instrumento comparador em qualquer população em estudo, não sendo específico para esta população. Respeitante à validade de critério, a mesma foi qualificada como boa devido ao tamanho da amostra em estudo. Por fim, a capacidade de resposta foi avaliada como fraca uma vez que o intervalo de tempo não foi descrito adequadamente.

O questionário do SPADI tem sido bastante utilizado na avaliação da funcionalidade e dor no ombro em estudos de efetividade de tratamentos. Num estudo para determinar a eficácia da proloterapia (terapia injetável) em utentes com doença crónica não traumática da coifa dos rotadores (tendinopatias, ruturas parciais e totais) foi utilizado o SPADI como instrumento de avaliação, por forma a monitorizar a evolução dos utentes (Lee, Kwack, Rah Woo, & Yoon, 2015). Num estudo comparativo entre duas intervenções (programa de exercícios em carga autónomo versus tratamento de fisioterapia) foi utilizado o SPADI por forma a avaliar as alterações na funcionalidade e dor no ombro em utentes com tendinopatia da coifa dos rotadores (Littlewood, Malliaras, Mawson, May, & Walters, 2014). Noutro estudo foi utilizado o SPADI para acompanhar as alterações na dor e funcionalidade de utentes com dor no ombro após terem sido intervencionados em cuidados primários, nos períodos de três semanas, três, seis e doze meses (Laslett, Steele, Hing, McNair, & Cadogan, 2014).

Assim, os resultados do nosso estudo evidenciaram que, de acordo com os critérios de avaliação das propriedades psicométricas, estas foram todas avaliadas com classificação positiva exceto a validade de critério e a capacidade de resposta. Na validade de critério a maioria das correlações existentes são inferiores a 0.7 e na capacidade de resposta pela inexistência de correlações entre as pontuações do SPADI e as pontuações dos instrumentos de comparação.

Relativamente à avaliação da metodologia do estudo referente ao questionário do SPADI (ANEXO II), a consistência interna apresentou-se fraca devido ao reduzido tamanho da amostra, à ausência da análise fatorial sem referência a outro estudo e ao facto de o ajuste estatístico a nível global não ter sido calculado. Relativamente à confiabilidade, a mesma foi fraca considerando que foi utilizada uma amostra diminuta. A validade de conteúdo evidenciou-se igualmente fraca, dado que não verificaram se os

itens em conjunto refletiam o constructo a ser medido, se estes se referiam a aspetos relevantes e se eram relevantes para a população em estudo uma vez que apenas foi realizada a análise de conteúdo da primeira questão de cada dimensão. Em relação ao teste de hipóteses, este classificou-se como fraco devido ao tamanho da amostra e pela ausência de informação acerca das propriedades de medida do instrumento de comparação. De igual forma, tanto a validade critério como a capacidade de resposta foram classificadas como fracas pois a amostra para ambas as propriedades foi reduzida. Esta classificação fraca do instrumento, em algumas das propriedades, deve-se ao número de indivíduos incluídos no estudo, sendo que participaram apenas 29 pessoas. Contudo, segundo os critérios da COSMIN, a classificação subiria para média na avaliação da metodologia da confiabilidade, validade de critério e capacidade de resposta se o número de participantes fosse superior a 30.

O questionário SRQ-PT avalia a sintomatologia e a funcionalidade do ombro. Desta forma foi realizado um estudo que se propôs comparar a eficácia de quatro questionários específicos do ombro na perceção da dor em cuidados de saúde primários, de entre os quais foi incluído o SRQ-PT. Neste estudo o SRQ-PT foi um dos questionários que apresentou maior capacidade de resposta (Paul et al., 2004). Um outro estudo recorreu à aplicação do SRQ-PT a fim de avaliar a eficácia de um programa de exercícios terapêuticos destinados a reduzir a dor e melhorar a função no ombro em trabalhadores de construção civil (Ludewig & Borstad, 2003).

No momento de avaliação das propriedades psicométricas, segundo os critérios de qualidade, a consistência interna e a validade de conteúdo foram classificadas como positivas, contudo a confiabilidade foi considerada indeterminada dado que nem o valor do CCI nem a correlação de *Pearson* foram calculados. Contudo esta propriedade psicométrica foi calculada utilizando o coeficiente de correlação de *Spearman*, para o qual o valor foi de 0.90.

Em relação à avaliação da metodologia do artigo referente ao questionário do SRQ-PT (ANEXO II), a consistência interna revelou ser fraca uma vez que a análise fatorial não foi realizada, não sendo referido outro estudo e o ajuste estatístico a nível global não foi calculado. Relativamente à confiabilidade, esta foi considerada fraca dado que as condições de teste não foram similares. Em contrapartida, a validade de conteúdo foi

classificada como excelente, não apresentando falhas a este nível. De forma a colmatar as lacunas existentes foi estabelecido contacto com o autor correspondente do artigo, no qual foi obtida resposta, no entanto não acrescentou novos dados à nossa revisão.

Embora o questionário UEFI avalie a funcionalidade de todo o membro superior, este tem sido utilizado em alguns estudos especificamente na avaliação da funcionalidade do ombro. Desta forma, o UEFI foi utilizado para avaliar a funcionalidade do ombro num estudo em que foi efetuada uma comparação entre duas intervenções (aplicação de tala com exercícios versus exercícios) em indivíduos com queimaduras axilares (Kolmus, Holland, Byrne, & Cleland, 2012). Noutro estudo, no qual os autores compararam a capacidade de resposta de diversos instrumentos de autorresposta em utentes com patologias da coifa dos rotadores, verificaram que o UEFI é o que apresenta melhor sensibilidade à mudança (capacidade de resposta) (Razmjou, Bean, van Osnabrugge, MacDermid, & Holtby, 2006).

Na nossa revisão, na avaliação das propriedades psicométricas do questionário do UEFI, segundo os critérios de qualidade, estas foram classificadas todas como positivas.

Na avaliação da metodologia do estudo referente ao UEFI (ANEXO II), a consistência interna revelou ser fraca uma vez que o tamanho da amostra é reduzido, a análise fatorial não foi realizada, não mencionando outro estudo e não foi calculado um ajuste estatístico global. Outra das propriedades considerada como fraca foi a confiabilidade, devido ao tamanho da amostra inadequado. Em oposição, a validade de conteúdo apresentou um grau de excelência, não revelando quaisquer falhas a esse nível.

O WUSPI tem como finalidade medir a intensidade da dor no ombro em utilizadores de cadeira de rodas durante a execução de atividades da vida diária. Desta forma, o objetivo de um estudo foi identificar a relação de dor no ombro com qualidade de vida, atividade física e atividades sociais em pessoas com paraplegia. Para tal utilizaram como métodos de avaliação diferentes tipos de instrumentos um dos quais o WUSPI (Gutierrez, Thompson, Kemp, & Mulroy, 2007). Num estudo realizado em utilizadores de cadeiras de rodas como forma de verificar a efetividade de exercícios de fortalecimento e alongamento, foi utilizado o WUSPI como medida de avaliação da dor no ombro (Nawoczenski, Ritter-Soronen, Wilson, Howe, & Ludewig, 2006). Noutro estudo para testar a eficácia do exercício autónomo na dor no ombro foram utilizados

como forma de medição o WUSPI, o DASH e o SRQ (Straaten, Cloud, Morrow, Ludewig, & Zhao, 2014).

Nos resultados da nossa revisão, as propriedades psicométricas do questionário do WUSPI foram todas consideradas como positivas à exceção do teste de hipóteses e da validade de critério que se classificaram como negativo. O primeiro foi considerado como negativo dada a existência de correlações não significativas em alguns parâmetros comparadores, não confirmando pelo menos 75% das hipóteses formuladas inicialmente. Em relação à validade de critério, foi classificada como negativa devido ao facto de não existirem correlações significativas entre os valores obtidos no WUSPI e no SF-36.

Na avaliação da metodologia do estudo referente ao WUSPI (ANEXO II), a consistência interna foi classificada como fraca devido ao tamanho da amostra diminuto, pelo facto de a análise fatorial não ter sido realizada, não referindo outro estudo e pelo ajuste estatístico global não calculado. Por outro lado, a validade de conteúdo foi considerada excelente, não apresentando quaisquer falhas. A confiabilidade, o teste de hipóteses, a validade de critério e a capacidade de resposta foram classificadas como fracas devido ao tamanho da amostra. Neste estudo, à semelhança do estudo do SPADI, teria sido classificado com uma melhor qualidade metodológica se a amostra utilizada tivesse incluído mais indivíduos. No estudo foram utilizadas subamostras com diferente número de indivíduos, sendo todas inferiores a 30 (12 indivíduos para avaliação da validade de conteúdo; 15 indivíduos para a consistência interna e confiabilidade; 28 indivíduos para o teste de hipóteses, validade de critério e capacidade de resposta). Caso a amostra tivesse mais de 30 indivíduos, segundo a classificação COSMIN, a qualidade metodológica para a confiabilidade, teste de hipóteses, validade de critério e capacidade de resposta teria sido avaliada como média.

De um modo geral, o tamanho da amostra incluída revelou ser reduzida, diminuindo frequentemente a qualidade metodológica dos estudos, sendo esta uma lacuna transversal a todos os instrumentos. Além disso, verificou-se ainda que nenhum dos estudos analisou o erro de medida como propriedade psicométrica, existindo, assim, uma falha na medição das mudanças que não estejam relacionadas com o constructo a

ser medido, não sendo possível medir o erro sistemático e aleatório das pontuações. A validade estrutural também não foi analisada em nenhum dos estudos, não sendo possível estimar o grau para o qual as pontuações do instrumento refletem adequadamente a dimensão do constructo a ser medido. Outro dos aspetos a salientar reside nas falhas metodológicas dos estudos dado que, por vezes, não foram realizadas as análises estatísticas adequadas ou estas não se encontravam descritas, o que levou a lacunas na apresentação e descrição dos dados obtidos em alguns estudos.

5.1 OUTRAS CONSIDERAÇÕES

Numa revisão sistemática realizada por Bot et al. que reuniu 16 questionários, a maioria dos estudos incluídos referem-se ao DASH e ao SPADI, escalas também incluídas na nossa revisão. Esta revisão da literatura propôs-se a estudar a qualidade psicométrica dos questionários incluídos, de forma a identificar todos os instrumentos que avaliem a incapacidade e função física do ombro. Segundo os resultados deste estudo, o DASH é o questionário que apresenta melhor classificação aquando da avaliação das propriedades de medida (Bot et al., 2004). Um outro estudo de Desai et al., recolheu 5 questionários comuns de autorresposta relativos ao ombro e avaliaram-nos de forma crítica no que toca ao seu desenvolvimento, validade, confiabilidade, capacidade de resposta e aplicação clínica. Dos 5 instrumentos recolhidos, incluíram-se o DASH, SPADI e SRQ. As conclusões deste estudo evidenciaram de igual forma que o DASH é o questionário que apresenta melhor classificação após a análise das propriedades psicométricas (Desai, Dramis, & Hearnden, 2010). Similarmente, numa revisão sistemática de Roy et al., após a análise das propriedades de medida de 4 instrumentos de incapacidade do ombro, entre os quais foram incluídos o DASH e o SPADI, os resultados evidenciaram que todos os questionários são aceitáveis para uso clínico (Roy et al., 2009). Segundo o artigo de Huang et al., o DASH e o SPADI mostraram ter um nível de evidência bom/moderado positivo para todas as propriedades psicométricas. Os mesmos questionários demonstraram ter boa evidência a suportar a consistência interna, a confiabilidade, a validade de conteúdo, o teste de hipóteses e a capacidade de resposta (Huang et al., 2015).

Numa revisão sistemática realizada por Puga et al. analisou o processo de tradução/adaptação cultural e das propriedades de medida de questionários que avaliam

a dor e as disfunções no ombro, traduzidos/adaptados para o português do brasil. Neste estudo, à semelhança da presente revisão sistemática, foram incluídos o DASH e o SPADI e concluíram que os instrumentos incluídos evidenciaram um processo de tradução de boa qualidade, contudo, as propriedades psicométricas não foram testadas sob as melhores condições (obtendo uma classificação metodológica fraca) (Puga et al., 2012). No entanto, nas outras revisões sistemáticas já referidas, os estudos incluídos demonstraram processos de validação mais rigorosos, sendo assim possível aferir uma classificação a cada um dos questionários estudados, dos quais o DASH foi o que apresentou melhores resultados.

5.2 LIMITAÇÕES E PONTOS FORTES

Quanto às limitações do presente estudo destaca-se o número reduzido de estudos, dado que para cada instrumento avaliado foi apenas encontrado um estudo referente à análise das suas propriedades psicométricas. Para além disso poucos são os estudos que foram submetidos a revisão e publicados em jornais ou revistas científicas, sendo que apenas os artigos referentes à avaliação das propriedades do DASH e SRQ-PT se encontram publicados. Os restantes estudos são trabalhos académicos, sendo que apenas o NULI-20 se encontra disponibilizado *online*. Para além disso, denotou-se alguma indisponibilidade da comunidade em partilhar estudos e informações, sendo estes de difícil acesso e com lacunas de informação. Por fim, salienta-se a falta de um *expert* no grupo aquando das discórdias entre o mesmo. Desta forma consideramos que estas implicações poderão de alguma forma ter comprometido a qualidade do presente estudo.

Não obstante destes factos, denotam-se ainda alguns pontos fortes. Neste aspeto, temos o nosso estudo como algo nunca antes realizado em Portugal, fornecendo um instrumento muito útil aos profissionais de saúde aquando da escolha de instrumentos de avaliação, permitindo um melhor cuidado centrado no utente. Outro dos pontos fortes foi o uso da nova nomenclatura baseada em critérios internacionais (proposta pelo grupo COSMIN), o uso de critérios de qualidade padronizados utilizados na elaboração de revisões sistemáticas de propriedades psicométricas, bem como a utilização da ferramenta PRISMA (para guiar a nossa revisão sistemática), conferindo-lhe uma maior qualidade metodológica.

5.3 CONCLUSÕES

Embora a maioria dos instrumentos apresente boas características psicométricas, a metodologia dos estudos incluídos apresenta algumas lacunas, classificando-se frequentemente como fraca. Desta forma, não é viável considerar os resultados dos estudos como totalmente fidedignos, descartando a possibilidade de inferir qual o questionário mais apropriado à prática clínica dos profissionais de saúde.

Neste estudo foram incluídos os instrumentos NULI-20 e WUSPI que, embora não sejam destinados à população geral, foram analisados por forma a fornecer uma base científica acerca da qualidade do instrumento. Uma vez que se destinam a populações específicas, e não havendo outros instrumentos que avaliem o mesmo constructo, importa conhecer a qualidade destes para que, aquando da sua aplicação, o profissional de saúde possa estar consciente das limitações das conclusões retiradas.

Aquando da escolha de um instrumento de autorresposta, os profissionais de saúde deverão estar atentos a estas considerações, primando pelo rigor na sua escolha atentando à qualidade metodológica dos estudos de validação das escalas e não apenas à qualidade das propriedades psicométricas. São necessários mais estudos de tradução e validação de instrumentos de autorresposta em Portugal que utilizem a terminologia aceite internacionalmente (COSMIN) e com qualidade metodológica mais elevada.

5.4 RECOMENDAÇÕES PARA O FUTURO

Recomenda-se a realização de novos estudos de validação para os instrumentos já traduzidos e validados para a população portuguesa com melhor metodologia, visto que os existentes apresentam qualidade metodológica fraca. Sugere-se ainda a realização de estudos semelhantes para outros constructos que possam incluir outro tipo de instrumentos.

6.CONCLUSÃO

Após todo este processo de aprendizagem, podemos concluir que os objetivos desta Unidade Curricular foram atingidos. Apesar das diversas dificuldades com que nos deparámos, nomeadamente ao nível de todo o processo de pesquisa e interpretação de dados que nos conduziria aos resultados e conclusões, consideramos que foi um processo desafiante e enriquecedor. Este percurso contribuiu para o nosso crescimento não só enquanto estudantes em fase de término de licenciatura, como a nível pessoal.

Assim, com a conclusão deste estudo foi possível perceber as lacunas existentes nesta área em Portugal, funcionando como um despertar enquanto profissionais de saúde para uma prática baseada na evidência. Esta monografia permitiu-nos atentar ao tipo de conduta na área da saúde, na qual os profissionais utilizam diversos instrumentos, muitas vezes sem questionar a sua veracidade. Assim é importante que esta revisão não seja meramente um trabalho académico, mas que leve a comunidade a refletir, a partilhar conhecimento, contribuindo para o aumento da qualidade dos cuidados de saúde prestados à comunidade. Para esse efeito foi criada uma versão artigo do nosso trabalho que será posteriormente submetido à Revista Ata Médica Portuguesa para publicação (ver Anexo III).

Num contexto futurístico, esperamos que o nosso trabalho possa ser reconhecido por todos os profissionais de saúde na sua prática clínica, como uma boa ferramenta aquando da seleção do melhor instrumento de medida no processo avaliativo de um utente com alterações na funcionalidade do CAO.

7.BIBLIOGRAFIA

APED. (2010a). *Dor no Ombro*.

APED. (2010b). Epidemiologia da Dor Musculoesquelética.

Bartels, E. M. (2013). How to perform a systematic search. *Best Practice & Research Clinical Rheumatology*, *27*(2), 295–306.

Bot, S. D. M., Terwee, C. B., van der Windt, D. A. W. M., Bouter, L. M., Dekker, J., & de Vet, H. C. W. (2004). Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Annals of the Rheumatic Diseases*, *63*(4), 335–41.

Bron, C., de Gast, A., Dommerholt, J., Stegenga, B., Wensing, M., & Oostendorp, R. a B. (2011). Treatment of myofascial trigger points in patients with chronic shoulder pain: a randomized, controlled trial. *BMC Medicine*, *9*(1), 8–22.

Clara, F. (2001). *Validação intercultural do Wheelchair User's Shoulder Pain Index:Monografia*.

Davies, H. T. O., & Crombie, I. K. (2001). What is a systematic review ? *Hayward Medical Communications*, *1*(5), 1–6.

Desai, A. S., Dramis, A., & Hearnden, A. J. (2010). Critical appraisal of subjective outcome measures used in the assessment of shoulder disability. *Annals of the Royal College of Surgeons of England*, 9–13.

DGS. (2008). Lesões Músculo-Esqueléticas Relacionadas com o Trabalho. *Gráfica Maiadouro, S.A.*, 1–24.

Duarte, A. (2002). *Validação intercultural do Shoulder Pain and Disability Index-SPADI:Monografia*.

Evans, D., & Pearson, A. (2001). Systematic reviews of qualitative research. *Clinical Effectiveness in Nursing*, *5*(3), 111–119.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Sloan, J. A. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health*, *10*(SUPPL. 2), 94–105.

Guerreiro, J. A., Proença, I., Moura, N., & Cartucho, A. (2011). Adaptação transcultural do Shoulder Rating Questionnaires para a Língua portugues (SRQ-PT): Tradução; Validação; Análise da consistência interna e replicabilidade. *Ifisionline*, *1*(2), 5–18.

Gutierrez, D. D., Thompson, L., Kemp, B., & Mulroy, S. J. (2007). The relationship of shoulder pain intensity to quality of life, physical activity, and community participation in persons with paraplegia. *The Journal of Spinal Cord Medicine*, *30*(3), 251–255.

Hanratty, C. E., McVeigh, J. G., Kerr, D. P., Basford, J. R., Finch, M. B., Pendleton, A., & Sim, J. (2012). The Effectiveness of Physiotherapy Exercises in Subacromial Impingement Syndrome: A Systematic Review and Meta-Analysis. *Seminars in Arthritis and Rheumatism*, *42*(3), 297–316.

Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health and Mental Health Services Research*, *34*(3), 283–291.

Huang, H., Grant, J. A., Miller, B. S., Mirza, F. M., & Gagnier, J. J. (2015). A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease. *The American Journal of Sports Medicine*, *43*(10), 2572–2582.

Ibrahim, M., Donatelli, R., Hellman, M., & Echternach, J. (2013). Efficacy of a static progressive stretch device as an adjunct to physical therapy in treating adhesive capsulitis of the shoulder: A prospective, randomised study. *Physiotherapy (United Kingdom)*, *100*(3), 228–234.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, *96*, 118–121.

Kolmus, A. M., Holland, A. E., Byrne, M. J., & Cleland, H. J. (2012). The effects of splinting on shoulder function in adult burns. *Burns*, *38*(5), 638–644.

Kyte, D. G., Calvert, M., van der Wees, P. J., ten Hove, R., Tolan, S., & Hill, J. C. (2015).

An introduction to patient-reported outcome measures (PROMs) in physiotherapy. *Physiotherapy (United Kingdom)*, *101*(2), 119–125.

Laslett, M., Steele, M., Hing, W., McNair, P., & Cadogan, A. (2014). Shoulder pain patients in primary care - Part 1: Clinical outcomes over 12 months following standardized diagnostic workup, corticosteroid injections, and community-based care. *Journal of Rehabilitation Medicine*, *46*(9), 898–907.

Lee, D.-H., Kwack, K.-S., Rah Woo, U., & Yoon, S.-H. (2015). Prolotherapy for Refractory Rotator Cuff Disease: Retrospective Case-Control Study of 1-Year Follow-Up. *Archives of Physical Medicine and Rehabilitation*, *96*(11), 2027–2032.

Littlewood, C., Malliaras, P., Mawson, S., May, S., & Walters, S. J. (2014). Self-managed loaded exercise versus usual physiotherapy treatment for rotator cuff tendinopathy: A pilot randomised controlled trial. *Physiotherapy (United Kingdom)*, *100*(1), 54–60.

Ludewig, P., & Borstad, J. (2003). Effects of a home exercise programme on shoulder pain and functional status in construction workers. *Occupational and Environmental Medicine*, *60*(11), 841–849.

Luime, J., Koes, B., Hendriksen, I., Burdorf, A., Verhagen, A., Miedema, H., & Verhaar, J. (2004). Prevalence and incidence of shoulder pain in the general population; a systematic review. *Scandinavian Journal of Rheumatology*, *33*(2), 73–81.

Matias, S. et al. (2010). *Tradução e adaptação cultural do Neck and Upper Limb Index para a língua portuguesa:Tese de Mestrado*.

Melo, F. (2002). *Upper Extremity Functional Index- Adaptação cultural e linguística:Monografia*.

Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., … De Vet, H. C. W. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research*, *18*(3), 313–333.

Mokkink, L. B., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., … De Vet, H. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international

Delphi study. *Quality of Life Research*, *19*(4), 539–549.

Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., … de Vet, H. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745.

Nawoczenski, D. a, Ritter-Soronen, J. M., Wilson, C. M., Howe, B. a, & Ludewig, P. M. (2006). Clinical trial of exercise for shoulder pain in chronic spinal injury. *Physical Therapy*, *86*(12), 1604–1618.

Paternostro-Sluga, T., & Zöch, C. (2004). Konservative Therapie und Rehabilitation von Schultererkrankungen. *Schulterdiagnostik*, 597–603.

Paul, A., Lewis, M., Shadforth, M. F., Croft, P. R., Van Der Windt, D. a W. M., & Hay, E. M. (2004). A comparison of four shoulder-specific questionnaires in primary care. *Annals of the Rheumatic Diseases*, *63*(10), 1293–1299.

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, *52*(11), 1746–1753.

Puga, V. O., Lopes, A. D., & Costa, L. O. (2012). Assessment of cross-cultural adaptations and measurement properties of self-report outcome measures relevant to shoulder disability in Portuguese: a systematic review. *Revista Brasileira de Fisioterapia*, *16*(2), 85–93.

Razmjou, H., Bean, A., van Osnabrugge, V., MacDermid, J. C., & Holtby, R. (2006). Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. *Bmc Musculoskeletal Disorders*, *7*(1), 26–33.

Rodrigues, A., Santana, M., & Pinheira, V. (2014). Avaliação da efetividade de um protocolo de Exercícios de Relaxamento Muscular em músicos com alterações músculoesqueléticas. *Revista de Saúde Pública*, 33–103.

Roy, J., MacDermid, J., & Woodhouse, L. (2009). Measuring Shoulder Function : A Systematic Review of Four Questionnaires. *Arthritis & Rheumatism (Arthritis Care & Research)*, *61*(5), 623–632.

Santos, J., & Gonçalves, R. (2006). Adaptação e validação cultural da versão portuguesa do

Disabilities of the Arm Shoulder and Hand–DASH. *Revista Portuguesa de Ortopedia E Traumatologia*, *14*, 29–46.

Schellingerhout, J. M., Verhagen, A. P., Heymans, M. W., Koes, B. W., De Vet, H. C., & Terwee, C. B. (2012). Measurement properties of disease-specific questionnaires in patients with neck pain: A systematic review. *Quality of Life Research*, *21*(4), 659–670.

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury*, *42*(3), 236–240.

Silva, M. (2006). Resultados de Medida. *Essfisionline*, *2*, 59–72.

Straaten, M., Cloud, B., Morrow, M., Ludewig, P., & Zhao, K. (2014). Effectiveness of Home Exercise on Pain, Function, and Strength of Manual Wheelchair Users With Spinal Cord Injury: A High-Dose Shoulder Program With Telerehabilitation. *Archives of Physical Medicine and Rehabilitation*, *95*(10), 1810–1817.

Terwee, C. (2011a). COSMIN checklist with 4-point scale.

Terwee, C. (2011b). Protocol for systematic reviews of measurement properties. *Measurement in Medicine*.

Terwee, C. B., Bot, S. D. M., Boer, M. R. De, Windt, A. W. M. Van Der, Knol, D. L., Dekker, J., … Vet, H. C. W. de. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42.

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & Vet, H. C. W. de. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist. *Quality of Life Research*, 651–657.

Terwee, C., Prinsen, C., Ricci Garotti, M., Suman, A., de Vet, H., & Mokkink, L. (2016). The quality of systematic reviews of health-related outcome measurement instruments. *Quality of Life Research*, *25*(4), 767–779.

van Tulder, M., Furlan, A., Bombardier, C., & Bouter, L. (2003). Updated method guidelines for systematic reviews in the cochrane collaboration Back Review Group. *Spine*, *28*(12), 1290–1299.

Winter, A., Heijden, G., Scholten, R., Windt, D., & Bouter, L. (2007). The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study. *Journal of Clinical Epidemiology*, *60*(11), 1156–1163.

ANEXOS

ANEXO I

Revisão da Literatura

REVISÃO DA LITERATURA

A revisão da literatura foi realizada em dois idiomas - inglês e português - pois embora a informação necessária seja relativa a Portugal esta poderia ter sido redigida na língua inglesa. Foi efetuada nas seguintes bases de dados: PubMed; PEDro; Google Académico; B-On e Repositório Científico de Acesso Aberto de Portugal (RCAAP); com as palavras-chave: "Assessment instrument"; "Outcome measure"; "PROM"; "Assessment scale"; "Measurement scale"; "Evaluation"; "Psychometric properties"; "Reliability"; "Validity"; "Responsiveness"; "Shoulder injury"; "Upper limb"; "Portuguese"; "Portugal"; "Instrumento de avaliação"; "Instrumento de medida"; "Avaliação"; "Instrumentos de autorresposta"; "Propriedades psicométricas"; "Validade"; "Confiabilidade"; "Reprodutibilidade"; "Capacidade de Resposta"; "Ombro"; "Membro superior". Todos estes conceitos foram conjugados através dos termos "OR" e "AND" (Bartels, 2013). Quando os resultados de pesquisa eram demasiado elevados, foram lidos os 200 primeiros títulos que surgiram no motor de busca.

Tabela de termos em inglês combinações com "AND" e "OR"			
Assessment instrument	Pshychometric properties	Upper Limb	Portugal
Outcome measure	Reliability	Shoulder Injury	
PROM	Validity		
Assessment scale	Responsiveness		
Measurement scale			
Evaluation			

Tabela de termos em português combinações com “AND” e “OR”			
Instrumento de avaliação	Propriedades psicométricas	Ombro	Portugal
Instrumento de medida	Validade	Membro superior	
Avaliação	Confiabilidade		
Instrumentos de autorresposta	Reprodutibilidade		
	Capacidade de resposta		

“Assessment instrument AND Psychometric properties AND Upper Limb”; “Assessment instrument AND Reliability AND Upper Limb”; “Assessment instrument AND Validity AND Upper Limb”; “Assessment instrument AND Responsiveness AND Upper Limb”; “Assessment instrument AND Psychometric properties AND Shoulder Injury”; “Assessment instrument AND Reliability AND Shoulder Injury”; “Assessment instrument AND Validity AND Shoulder Injury”; “Assessment instrument AND Responsiveness AND Shoulder Injury”; “Assessment instrument AND Psychometric properties AND Upper Limb AND Portugal”; “Assessment instrument AND Reliability AND Upper Limb AND Portugal”; “Assessment instrument AND Validity AND Upper Limb AND Portugal”; “Assessment instrument AND Responsiveness AND Upper Limb AND Portugal”; “Assessment instrument AND Psychometric properties AND Shoulder Injury AND Portugal”; “Assessment instrument AND Reliability AND Shoulder Injury AND Portugal”; “Assessment instrument AND Validity AND Shoulder Injury AND Portugal”; “Assessment instrument AND Responsiveness AND Shoulder Injury AND Portugal”; “Assessment instrument OR Psychometric properties AND Upper Limb”; “Assessment instrument OR Reliability AND Upper Limb”; “Assessment instrument OR Validity AND Upper Limb”; “Assessment instrument OR Responsiveness AND Upper Limb”; “Assessment instrument OR Psychometric properties AND Shoulder Injury”; “Assessment instrument OR Reliability AND Shoulder Injury”; “Assessment instrument OR Validity AND Shoulder Injury”; “Assessment instrument OR Responsiveness AND Shoulder Injury”; “Assessment instrument OR Psychometric properties AND Upper Limb AND Portugal”; “Assessment instrument OR Reliability AND Upper Limb AND Portugal”; “Assessment instrument OR Validity AND Upper Limb AND Portugal”; “Assessment

instrument OR Responsiveness AND Upper Limb AND Portugal"; "Assessment instrument OR Psychometric properties AND Shoulder Injury AND Portugal"; "Assessment instrument OR Reliability AND Shoulder Injury AND Portugal"; "Assessment instrument OR Validity AND Shoulder Injury AND Portugal"; "Assessment instrument OR Responsiveness AND Shoulder Injury AND Portugal".

"Outcome measure AND Pshychometric properties AND Upper Limb AND Portugal"; "Outcome measure OR Pshychometric properties OR Upper Limb OR Portugal"; "Outcome measure AND Pshychometric properties AND Shoulder Injury AND Portugal"; "Pshychometric properties OR Shoulder Injury OR Portugal"; "Outcome measure AND Reliability AND Upper Limb AND Portugal ";"Outcome measure OR Reliability OR Upper Limb OR Portugal" ; "Outcome measure AND Reliability AND Shoulder Injury AND Portugal ";"Outcome measure OR Reliability OR Shoulder Injury OR Portugal "Outcome measure AND Validity AND Upper Limb AND Portugal"; "Outcome measure OR Validity OR Upper Limb OR Portugal"; "Outcome measure AND Validity AND Shoulder Injury AND Portugal"; "Outcome measure OR validity OR Shoulder Injury OR Portugal"; "Outcome measure AND Responsiveness AND Upper Limb AND Portugal"; "Outcome measure OR Responsiveness OR Upper Limb OR Portugal"; "Outcome measure AND Responsiveness AND Shoulder Injury AND Portugal"; "Outcome measure OR Responsiveness OR Shoulder Injury OR Portugal".

"PROM AND Pshychometric properties AND Upper Limb AND Portugal"; " PROM OR Pshychometric properties OR Upper Limb OR Portugal" PROM AND Pshychometric properties AND Shoulder Injury AND Portugal" PROM OR Pshychometric properties OR Shoulder Injury OR Portugal"; " PROM AND Reliability AND Upper Limb AND Portugal"; "PROM OR Reliability OR Upper Limb OR Portugal"; "PROM AND Reliability AND Shoulder Injury AND Portugal"; "PROM OR Reliability OR Shoulder Injury OR Portugal"; "PROM AND Validity AND Upper Limb AND Portugal"; "PROM OR Validity OR Upper Limb OR Portugal" PROM AND Validity AND Shoulder Injury AND Portugal"; "PROM OR validity OR Shoulder Injury OR Portugal"; "PROM AND Responsiveness AND Upper Limb AND Portugal"; "PROM OR Responsiveness OR Upper Limb OR Portugal"; "PROM AND Responsiveness AND Shoulder Injury AND Portugal"; "PROM OR Responsiveness OR Shoulder Injury OR Portugal".

"Assessment scale AND Psychometric properties AND Upper Limb"; "Assessment scale AND Reliability AND Upper Limb"; "Assessment scale AND Validity AND Upper

Limb"; "Assessment scale AND Responsiveness AND Upper Limb"; "Assessment scale AND Psychometric properties AND Shoulder Injury"; "Assessment scale AND Reliability AND Shoulder Injury"; "Assessment scale AND Validity AND Shoulder Injury"; "Assessment scale AND Responsiveness AND Shoulder Injury"; "Assessment scale AND Psychometric properties AND Upper Limb AND Portugal"; "Assessment scale AND Reliability AND Upper Limb AND Portugal"; "Assessment scale AND Validity AND Upper Limb AND Portugal"; "Assessment scale AND Responsiveness AND Upper Limb AND Portugal"; "Assessment scale AND Psychometric properties AND Shoulder Injury AND Portugal"; "Assessment scale AND Reliability AND Shoulder Injury AND Portugal"; "Assessment scale AND Validity AND Shoulder Injury AND Portugal"; "Assessment scale AND Responsiveness AND Shoulder Injury AND Portugal"; "Assessment scale OR Psychometric properties OR Upper Limb"; "Assessment scale OR Reliability OR Upper Limb"; "Assessment scale OR Validity OR Upper Limb"; "Assessment scale OR Responsiveness OR Upper Limb"; "Assessment scale OR Psychometric properties OR Shoulder Injury"; "Assessment scale OR Reliability OR Shoulder Injury"; "Assessment scale OR Validity OR Shoulder Injury"; "Assessment scale OR Responsiveness OR Shoulder Injury"; "Assessment scale OR Psychometric properties OR Upper Limb AND Portugal"; "Assessment instrument OR Reliability OR Upper Limb AND Portugal"; "Assessment scale OR Validity OR Upper Limb AND Portugal"; "Assessment scale OR Responsiveness OR Upper Limb AND Portugal"; "Assessment scale OR Psychometric properties OR Shoulder Injury AND Portugal"; "Assessment scale OR Reliability OR Shoulder Injury AND Portugal"; "Assessment scale OR Validity OR Shoulder Injury AND Portugal"; "Assessment scale OR Responsiveness OR Shoulder Injury AND Portugal".

"Measurement scale AND Psychometric properties AND Upper Limb"; "Measurement scale AND Reliability AND Upper Limb"; "Measurement scale AND Validity AND Upper Limb"; "Measurement scale AND Responsiveness AND Upper Limb"; "Measurement scale AND Psychometric properties AND Shoulder Injury"; "Measurement scale AND Reliability AND Shoulder Injury"; "Measurement scale AND Validity AND Shoulder Injury"; "Measurement scale AND Responsiveness AND Shoulder Injury"; "Measurement scale AND Psychometric properties AND Upper Limb AND Portugal"; "Measurement scale AND Reliability AND Upper Limb AND Portugal"; "Measurement scale AND Validity AND Upper Limb AND Portugal"; "Measurement scale AND Responsiveness AND Upper Limb AND Portugal"; "Measurement scale AND Psychometric properties AND Shoulder Injury

AND Portugal"; "Measurement scale AND Reliability AND Shoulder Injury AND Portugal"; "Measurement scale AND Validity AND Shoulder Injury AND Portugal"; "Measurement scale AND Responsiveness AND Shoulder Injury AND Portugal"; "Measurement scale OR Psychometric properties OR Upper Limb"; "Measurement scale OR Reliability OR Upper Limb"; "Measurement scale OR Validity OR Upper Limb"; "Measurement scale OR Responsiveness OR Upper Limb"; "Measurement scale OR Psychometric properties OR Shoulder Injury"; "Measurement scale OR Reliability OR Shoulder Injury"; "Measurement scale OR Validity OR Shoulder Injury"; "Measurement scale OR Responsiveness OR Shoulder Injury"; "Measurement scale OR Psychometric properties OR Upper Limb AND Portugal"; "Measurement instrument OR Reliability OR Upper Limb AND Portugal"; "Measurement scale OR Validity OR Upper Limb AND Portugal"; "Measurement scale OR Responsiveness OR Upper Limb AND Portugal"; "Measurement scale OR Psychometric properties OR Shoulder Injury AND Portugal"; "Measurement scale OR Reliability OR Shoulder Injury AND Portugal"; "Measurement scale OR Validity OR Shoulder Injury AND Portugal"; "Measurement scale OR Responsiveness OR Shoulder Injury AND Portugal".

"Evaluation AND Psychometric Properties AND Upper Limb AND Portugal"; "Evaluation OR Psychometric Properties OR Upper Limb AND Portugal"; "Evaluation AND Reliability AND Upper Limb AND Portugal"; "Evaluation OR Reliability OR Upper Limb AND Portugal"; "Evaluation AND Validity AND Upper Limb AND Portugal"; "Evaluation OR Validity OR Upper Limb AND Portugal"; "Evaluation AND Responsiveness AND Upper Limb AND Portugal"; "Evaluation OR Responsiveness OR Upper Limb AND Portugal"; "Evaluation AND Psychometric properties AND Shoulder Injury AND Portugal"; "Evaluation OR Psychometric properties OR Shoulder Injury AND Portugal"; "Evaluation AND Reliability AND Shoulder Injury AND Portugal"; "Evaluation OR Reliability OR Shoulder Injury AND Portugal"; "Evaluation AND Validity AND Shoulder Injury AND Portugal"; "Evaluation OR Validity OR Shoulder Injury AND Portugal"; "Evaluation AND Responsiveness AND Shoulder Injury AND Portugal"; "Evaluation OR Responsiveness OR Shoulder Injury AND Portugal"; "Evaluation AND Psychometric Properties AND Upper Limb AND Portuguese"; "Evaluation OR Psychometric Properties OR Upper Limb AND Portuguese"; "Evaluation AND Reliability AND Upper Limb AND Portuguese"; "Evaluation OR Reliability OR Upper Limb AND Portuguese"; "Evaluation AND Validity AND Upper Limb AND Portuguese"; "Evaluation OR Validity OR Upper Limb AND Portuguese"; "Evaluation AND Responsiveness AND Upper Limb AND Portuguese";

“Evaluation OR Responsiveness OR Upper Limb AND Portuguese”; “Evaluation AND Psychometric properties AND Shoulder Injury AND Portuguese”; “Evaluation OR Psychometric properties OR Shoulder Injury AND Portuguese”; “Evaluation AND Reliability AND Shoulder Injury AND Portuguese”; “Evaluation OR Reliability OR Shoulder Injury AND Portuguese”; “Evaluation AND Validity AND Shoulder Injury AND Portuguese”; “Evaluation OR Validity OR Shoulder Injury AND Portuguese”; “Evaluation AND Responsiveness AND Shoulder Injury AND Portuguese”; “Evaluation OR Responsiveness OR Shoulder Injury AND Portuguese.

“Instrumento de avaliação AND Propriedades Psicométricas AND Ombro AND Portugal”; “Instrumento de avaliação OR Propriedades Psicométricas OR Ombro AND Portugal”; “Instrumento de avaliação AND Validade AND Ombro AND Portugal”; “Instrumento de avaliação OR Validade OR Ombro AND Portugal”; “Instrumento de avaliação AND Confiabilidade AND Ombro AND Portugal”; “Instrumento de avaliação OR Confiabilidade OR Ombro AND Portugal”; “Instrumento de avaliação AND Reprodutibilidade AND Ombro AND Portugal”; “Instrumento de avaliação OR Reprodutibilidade OR Ombro AND Portugal”; “Instrumento de avaliação AND Capacidade de resposta AND Ombro AND Portugal”; “Instrumento de avaliação OR Capacidade de resposta OR Ombro AND Portugal”; “Instrumento de avaliação AND Propriedades Psicométricas AND Membro superior AND Portugal”; “Instrumento de avaliação OR Propriedades Psicométricas OR Membro superior AND Portugal”; “Instrumento de avaliação AND Validade AND Membro superior AND Portugal”; “Instrumento de avaliação OR Validade OR Membro superior AND Portugal”; “Instrumento de avaliação AND Confiabilidade AND Membro superior AND Portugal”; “Instrumento de avaliação OR Confiabilidade OR Membro superior AND Portugal”; “Instrumento de avaliação AND Reprodutibilidade AND Membro superior AND Portugal”; “Instrumento de avaliação OR Reprodutibilidade OR Membro superior AND Portugal”; “Instrumento de avaliação AND Capacidade de resposta AND Membro superior AND Portugal”; “Instrumento de avaliação OR Capacidade de resposta OR Membro superior AND Portugal”.

“Instrumento de medida AND Propriedades psicométricas AND Ombro”; “Instrumento de medida AND Validade AND Ombro”; “Instrumento de medida AND Confiabilidade AND Ombro”; “Instrumento de medida AND Reprodutibilidade AND Ombro”; “Instrumento de medida AND Capacidade de Resposta AND Ombro”; “Instrumento de medida AND Propriedades psicométricas AND Membro Superior”; “Instrumento de medida AND

Validade AND Membro Superior"; "Instrumento de medida AND Confiabilidade AND Membro Superior"; "Instrumento de medida AND Reprodutibilidade AND Membro Superior"; "Instrumento de medida AND Capacidade de Resposta AND Membro Superior"; "Instrumento de medida AND Propriedades psicométricas AND Ombro AND Portugal"; "Instrumento de medida AND Validade AND Ombro AND Portugal"; "Instrumento de medida AND Confiabilidade AND Ombro AND Portugal"; "Instrumento de medida AND Reprodutibilidade AND Ombro AND Portugal"; "Instrumento de medida AND Capacidade de Resposta AND Ombro AND Portugal"; "Instrumento de medida AND Propriedades psicométricas AND Membro Superior AND Portugal"; "Instrumento de medida AND Validade AND Membro Superior AND Portugal"; "Instrumento de medida AND Confiabilidade AND Membro Superior AND Portugal"; "Instrumento de medida AND Reprodutibilidade AND Membro Superior AND Portugal"; "Instrumento de medida AND Capacidade de Resposta AND Membro Superior AND Portugal"; "Instrumento de medida OR Propriedades psicométricas AND Ombro"; "Instrumento de medida OR Validade AND Ombro"; "Instrumento de medida OR Confiabilidade AND Ombro"; "Instrumento de medida OR Reprodutibilidade AND Ombro"; "Instrumento de medida OR Capacidade de Resposta AND Ombro"; "Instrumento de medida OR Propriedades psicométricas AND Membro Superior"; "Instrumento de medida OR Validade AND Membro Superior"; "Instrumento de medida OR Confiabilidade AND Membro Superior"; "Instrumento de medida OR Reprodutibilidade AND Membro Superior"; "Instrumento de medida OR Capacidade de Resposta AND Membro Superior"; "Instrumento de medida OR Propriedades psicométricas AND Ombro AND Portugal"; "Instrumento de medida OR Validade AND Ombro AND Portugal"; "Instrumento de medida OR Confiabilidade AND Ombro AND Portugal"; "Instrumento de medida OR Reprodutibilidade AND Ombro AND Portugal"; "Instrumento de medida OR Capacidade de Resposta AND Ombro AND Portugal"; "Instrumento de medida OR Propriedades psicométricas AND Membro Superior AND Portugal"; "Instrumento de medida OR Validade AND Membro Superior AND Portugal"; "Instrumento de medida OR Confiabilidade AND Membro Superior AND Portugal"; "Instrumento de medida OR Reprodutibilidade AND Membro Superior AND Portugal"; "Instrumento de medida OR Capacidade de Resposta AND Membro Superior AND Portugal".

"Avaliação AND Propriedades Psicométricas AND Ombro AND Portugal"; "Avaliação OR Propriedades Psicométricas OR Ombro AND Portugal"; "Avaliação AND Validade AND Ombro AND Portugal"; "Avaliação OR Validade OR Ombro AND Portugal"; "Avaliação

AND Confiabilidade AND Ombro AND Portugal"; "Avaliação OR Confiabilidade OR Ombro AND Portugal"; "Avaliação AND Reprodutibilidade AND Ombro AND Portugal"; "Avaliação OR Reprodutibilidade OR Ombro AND Portugal"; "Avaliação AND Capacidade de resposta AND Ombro AND Portugal"; "Avaliação OR Capacidade de resposta OR Ombro AND Portugal"; "Avaliação AND Propriedades Psicométricas AND Membro superior AND Portugal"; "Avaliação OR Propriedades Psicométricas OR Membro superior AND Portugal"; "Avaliação AND Validade AND Membro superior AND Portugal"; "Avaliação OR Validade OR Membro superior AND Portugal"; "Avaliação AND Confiabilidade AND Membro superior AND Portugal"; "Avaliação OR Confiabilidade OR Membro superior AND Portugal"; "Avaliação AND Reprodutibilidade AND Membro superior AND Portugal"; "Avaliação OR Reprodutibilidade OR Membro superior AND Portugal"; "Avaliação AND Capacidade de resposta AND Membro superior AND Portugal"; "Avaliação OR Capacidade de resposta OR Membro superior AND Portugal".

"Instrumento de autorresposta AND Propriedades psicométricas AND Ombro"; "Instrumento de autorresposta AND Validade AND Ombro"; "Instrumento de autorresposta AND Confiabilidade AND Ombro"; "Instrumento de autorresposta AND Reprodutibilidade AND Ombro"; "Instrumento de autorresposta AND Capacidade de Resposta AND Ombro"; "Instrumento de autorresposta AND Propriedades psicométricas AND Membro Superior"; "Instrumento de autorresposta AND Validade AND Membro Superior"; "Instrumento de autorresposta AND Confiabilidade AND Membro Superior"; "Instrumento de autorresposta AND Reprodutibilidade AND Membro Superior"; "Instrumento de autorresposta AND Capacidade de Resposta AND Membro Superior"; "Instrumento de autorresposta AND Propriedades psicométricas AND Ombro AND Portugal"; "Instrumento de autorresposta AND Validade AND Ombro AND Portugal"; "Instrumento de autorresposta AND Confiabilidade AND Ombro AND Portugal"; "Instrumento de autorresposta AND Reprodutibilidade AND Ombro AND Portugal"; "Instrumento de autorresposta AND Capacidade de Resposta AND Ombro AND Portugal"; "Instrumento de autorresposta AND Propriedades psicométricas AND Membro Superior AND Portugal"; "Instrumento de autorresposta AND Validade AND Membro Superior AND Portugal"; "Instrumento de autorresposta AND Confiabilidade AND Membro Superior AND Portugal"; "Instrumento de autorresposta AND Reprodutibilidade AND Membro Superior AND Portugal"; "Instrumento de autorresposta AND Capacidade de Resposta AND Membro Superior AND Portugal"; "Instrumento de autorresposta OR Propriedades psicométricas AND Ombro"; "Instrumento de autorresposta OR Validade AND Ombro"; "Instrumento de autorresposta

OR Confiabilidade AND Ombro"; "Instrumento de autorresposta OR Reprodutibilidade AND Ombro"; "Instrumento de autorresposta OR Capacidade de Resposta AND Ombro"; "Instrumento de autorresposta OR Propriedades psicométricas AND Membro Superior"; "Instrumento de autorresposta OR Validade AND Membro Superior"; "Instrumento de autorresposta OR Confiabilidade AND Membro Superior"; "Instrumento de autorresposta OR Reprodutibilidade AND Membro Superior"; "Instrumento de autorresposta OR Capacidade de Resposta AND Membro Superior"; "Instrumento de autorresposta OR Propriedades psicométricas AND Ombro AND Portugal"; "Instrumento de autorresposta OR Validade AND Ombro AND Portugal"; "Instrumento de autorresposta OR Confiabilidade AND Ombro AND Portugal"; "Instrumento de autorresposta OR Reprodutibilidade AND Ombro AND Portugal"; "Instrumento de autorresposta OR Capacidade de Resposta AND Ombro AND Portugal"; "Instrumento de autorresposta OR Propriedades psicométricas AND Membro Superior AND Portugal"; "Instrumento de autorresposta OR Validade AND Membro Superior AND Portugal"; "Instrumento de autorresposta OR Confiabilidade AND Membro Superior AND Portugal"; "Instrumento de autorresposta OR Reprodutibilidade AND Membro Superior AND Portugal"; "Instrumento de autorresposta OR Capacidade de Resposta AND Membro Superior AND Portugal".

Desta pesquisa foram obtidos 2 resultados: Adaptação Transcultural do *Shoulder Rating Questionnaire* para a Língua Portuguesa (SRQ-PT): Tradução; Validação; Análise da Consistência Interna e Replicabilidade, bem como a Tradução e Adaptação Cultural do *Neck and Upper Limb Index* para a Língua Portuguesa. Os restantes artigos: Adaptação e Validação Cultural da Versão Portuguesa do *Disabilities of the Arm Shoulder and Hand* – DASH; *Upper Extremity Functional Index* (UEFI) – Adaptação Cultural e Linguística; Validação Intercultural do *Shoulder Pain and Disability Index* – SPADI e Validação Intercultural do *Weelchair User´s Shoulder Pain Index* – WUSPI foram obtidos através do contacto com as escolas que lecionam Fisioterapia em Portugal, Centro de Estudos e Investigação em Saúde da Universidade de Coimbra (CEISUC) e Escola Superior de Tecnologias da Saúde de Coimbra.

ANEXO II

Avaliação metodológica segundo a COSMIN *checklist*

AVALIAÇÃO METODOLÓGICA SEGUNDO A COSMIN *CHECKLIST*

Disabilities of the Arm Shoulder and Hand (DASH)

Step 1. Evaluated measurement properties in the article

Fraco	Internal consistency	Box A
Médio	Reliability	Box B
	Measurement error	Box C
	Content validity	Box D
	Structural validity	Box E
Fraco	Hypotheses testing	Box F
Fraco	Cross-cultural validity	Box G
Bom	Criterion validity	Box H
	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Sim			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items

7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
9	for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10	for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11	for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

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

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box F. Hypotheses testing

		excellent	good	fair	Poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)

9	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods					
10	Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity					
		excellent	**good**	**fair**	**poor**
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods			
14	for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed	Multiple-group confirmatory factor analysis NOT performed
15	for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed	DIF between language groups NOT assessed

Box H. Criterion validity

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

5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated		Correlations or AUC NOT calculated
7	for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated		Sensitivity and specificity NOT calculated

Neck and Upper Limb Index (NULI-20)

Step 1. Evaluated measurement properties in the article

Fraco	Internal consistency	Box A
Medio	Reliability	Box B
	Measurement error	Box C
Excelente	Content validity	Box D
	Structural validity	Box E
Médio	Hypotheses testing	Box F
Fraco	Cross-cultural validity	Box G
Bom	Criterion validity	Box H
Fraco	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Não			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items

7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
9	for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10	for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11	for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

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

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate

9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)					
		excellent	**good**	**fair**	**poor**
General requirements					
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box F. Hypotheses testing				
	excellent	**good**	**fair**	**Poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5 Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6 Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods			
14 for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed		Multiple-group confirmatory factor analysis NOT performed
15 for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed		DIF between language groups NOT assessed

Box H. Criterion validity				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'

5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
6 for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
7 for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Box I. Responsiveness				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described

6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
Design requirements for hypotheses testing					
	For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11	Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12	Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods					
14	Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Design requirement for comparison to a gold standard					
	For constructs for which a gold standard was available:				
15	Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Shoulder Pain And Disability Index (SPADI)

Step 1. Evaluated measurement properties in the article

Fraco	Internal consistency	Box A
Fraco	Reliability	Box B
	Measurement error	Box C
Fraco	Content validity	Box D
	Structural validity	Box E
Fraco	Hypotheses testing	Box F
	Cross-cultural validity	Box G
Fraco	Criterion validity	Box H
Fraco	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Sim			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

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

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)

		excellent	good	fair	poor
General requirements					
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box F. Hypotheses testing

		excellent	good	fair	Poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)

9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box H. Criterion validity				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'

5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
6 for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
7 for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Box I. Responsiveness				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5 Was the time interval stated?	Time interval adequately described			Time interval NOT described

6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
	Design requirements for hypotheses testing				
	For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11	Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12	Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
	Statistical methods				
14	Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Design requirement for comparison to a gold standard					
	For constructs for which a gold standard was available:				
15	Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Shoulder Rating Questionnaire (SRQ-PT)

Fraco	Internal consistency	Box A
Fraco	Reliability	Box B
	Measurement error	Box C
Excelente	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
Fraco	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Sim			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

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

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)

		excellent	good	fair	poor
General requirements					
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box G. Cross-cultural validity

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods			
14	for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed	Multiple-group confirmatory factor analysis NOT performed
15	for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed	DIF between language groups NOT assessed

Upper Extremity Functional Index (UEFI)

Fraco	Internal consistency	Box A
Fraco	Reliability	Box B
	Measurement error	Box C
Excelente	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
Fraco	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Sim			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

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

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Were at least two measurements available?	At least two measurements			Only one measurement
5	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6	Was the time interval stated?	Time interval stated		Time interval NOT stated	
7	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13	for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)

		excellent	good	fair	poor
General requirements					
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box G. Cross-cultural validity

		excellent	good	fair	poor
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods			
14	for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed	Multiple-group confirmatory factor analysis NOT performed
15	for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed	DIF between language groups NOT assessed

Weelchair User's Shoulder Pain Index (WUSPI)

Fraco	Internal consistency	Box A
Fraco	Reliability	Box B
	Measurement error	Box C
Excelente	Content validity	Box D
	Structural validity	Box E
Fraco	Hypotheses testing	Box F
Fraco	Cross-cultural validity	Box G
Fraco	Criterion validity	Box H
Fraco	Responsiveness	Box I

Box A. Internal consistency				
	excellent	**good**	**fair**	**poor**
1 Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Sim			
Design requirements				
2 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4 Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
5 Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6 Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6-7* #items but <100	5* #items but <100	<5* #items
7 Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
9 for Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10 for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11 for IRT: Was a goodness of fit statistic at a global level calculated? E.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study.

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4 Were at least two measurements available?	At least two measurements			Only one measurement
5 Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent
6 Was the time interval stated?	Time interval stated		Time interval NOT stated	
7 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9 Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11 for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12 for dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13 for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14 for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box D. Content validity (including face validity)				
	excellent	**good**	**fair**	**poor**
General requirements				
1 Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2 Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size (≥10)	Assessed if all items are relevant for the study population in moderate sample size (5-9)	Assessed if all items are relevant for the study population in small sample size (<5)	NOT assessed if all items are relevant for the study population OR target population not involved
3 Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4 Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Box F. Hypotheses testing				
	excellent	**good**	**fair**	**Poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3 Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
4 Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5 Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6 Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7 for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8 for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)

9 Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods				
10 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box G. Cross-cultural validity				
	excellent	**good**	**fair**	**poor**
Design requirements				
1 Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2 Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

3	Was the sample size included in the analysis adequate?	CTT: 7* #items and ≥100 IRT: ≥200 per group	CTT: 5* #items and ≥100 OR 5-7* #items but <100 IRT: ≥200 in 1 group and 100-199 in 1 group	CTT: 5* #items but <100 IRT: 100-199 per group	CTT: <5* #items IRT: (<100 in 1 or both groups
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	Both source language and target language described			Source language NOT known
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	Expertise of the translators described with respect to disease, construct, and language	Expertise of the translators with respect to disease or construct poor or not described	Expertise of the translators with respect to language not described	
6	Did the translators work independently from each other?	Translators worked independent	Assumable that the translators worked independent	Unclear whether translators worked independent	Translators worked NOT independent
7	Were items translated forward and backward?	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation
8	Was there an adequate description of how differences between the original and translated versions were resolved?	Adequate description of how differences between translators were resolved	Poorly or NOT described how differences between translators were resolved		

9	Was the translation reviewed by a committee (e.g. original developers)?	Translation reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation NOT reviewed by (such) a committee		
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	Translated instrument pre-tested in the target population	Translated instrument pre-tested, but unclear if this was done in the target population	Translated instrument pre-tested, but NOT in the target population	Translated instrument NOT pre-tested
11	Was the sample used in the pre-test adequately described?	Sample used in the pre-test adequately described		Sample used in the pre-test NOT (adequately) described	
12	Were the samples similar for all characteristics except language and/or cultural background?	Shown that samples were similar for all characteristics except language /culture	Stated (but not shown) that samples were similar for all characteristics except language /culture	Unclear whether samples were similar for all characteristics except language /culture	Samples were NOT similar for all characteristics except language /culture
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study

Statistical methods			
14	for CTT: Was confirmatory factor analysis performed?	Multiple-group confirmatory factor analysis performed	Multiple-group confirmatory factor analysis NOT performed
15	for IRT: Was differential item function (DIF) between language groups assessed?	DIF between language groups assessed	DIF between language groups NOT assessed

Box H. Criterion validity					
		excellent	**good**	**fair**	**poor**
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'

5	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
7	for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Box I. Responsiveness					
		excellent	**good**	**fair**	**poor**
Design requirements					
1	Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2	Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3	Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
4	Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5	Was the time interval stated?	Time interval adequately described			Time interval NOT described

6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
Design requirements for hypotheses testing					
	For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

11	Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12	Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Statistical methods					
14	Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Design requirement for comparison to a gold standard					
	For constructs for which a gold standard was available:				
15	Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
16	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods					
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

ANEXO III

Versão pré-final do artigo para publicação

Avaliação da Qualidade dos Instrumentos de Medida nas Disfunções do Complexo Articular do Ombro – Revisão Sistemática

Quality Assessment of Measurement Instruments in Disorders of the Shoulder Complex - Systematic Review

Jéssica Silveira[1], Carina Vieira[1,3], Ana Freitas[1], Joana Palaio[1], Nuno Morais[2]

1- Estudante. Escola Superior de Saúde. Instituto Politécnico de Leiria. Leiria. Portugal.

2- Fisioterapeuta. Docente. Escola Superior de Saúde. Instituto Politécnico de Leiria. Leiria. Portugal.

3- Autor correspondente: Carina Vieira.

Castelo da Gracieira n31, 3100-083 Albergaria dos Doze, Pombal, Leiria

carinanevesvieira@gmail.com

Avaliação da Qualidade dos Instrumentos de Medida nas Disfunções do Complexo Articular do Ombro

Avaliação da Qualidade dos Instrumentos de Medida nas Disfunções do Complexo Articular do Ombro – Revisão Sistemática

RESUMO

Introdução: Dada a prevalência de disfunções no Complexo Articular do Ombro, os registos de avaliação segundo a perspetiva do utente constituem ferramentas úteis na seleção das estratégias de intervenção. A escolha do instrumento adequado deve-se basear em grande parte na força das suas propriedades psicométricas, contudo não existem estudos que analisem sistematicamente a qualidade destas medidas. Propomo-nos analisar sistematicamente estudos referentes às propriedades psicométricas de instrumentos de autorresposta nas disfunções no Complexo Articular do Ombro.

Materiais e Métodos: Revisão da literatura em inglês/português, nas bases de dados: PubMed, PEDro, Google Académico, B-On e RCAAP. Foram analisados artigos realizados até 2015. A qualidade metodológica e as propriedades psicométricas foram avaliadas e resumidas através de dois critérios padronizados, seguindo a ideologia COSMIN.

Resultados: Neste estudo foram incluídos 6 artigos. O *Disabilities of the Arm Shoulder and Hand* e o *Neck and Upper Limb Index* demonstram boas propriedades psicométricas e uma metodologia média; o *Shoulder Pain And Disability Index* e o *Weelchair User's Shoulder Pain Index* exibem boas propriedades psicométricas e qualidade metodológica fraca; tanto no *Shoulder Rating Questionnaire* como no *Upper Extremity Functional Index* não foram avaliadas propriedades de medida relevantes, contudo as analisadas apresentam boas propriedades psicométricas e uma metodologia fraca.

Discussão e conclusão: Devido às falhas na metodologia dos estudos incluídos, não é possível inferir qual o questionário mais apropriado à prática clínica. São necessários mais estudos de validação de instrumentos de autorresposta com melhor qualidade metodológica.

Palavras-chave: Complexo Articular do Ombro, Propriedades Psicométricas, Instrumentos de Autorresposta, COSMIN.

ABSTRACT

Background: Given the prevalence of shoulder complex dysfunctions, patient reported outcome measures are useful tools for choosing intervention strategies. The instrument must be selected mainly according the strength of its psychometric properties. However, there aren't any studies that systematically analyze the quality of these measures. We propose to systematically analyze studies referring the psychometric properties of patient reported outcome measures in the shoulder complex.

Material and Methods: Literature searches were performed in english/portuguese languages in the following databases: PubMed, PEDro, Google Scholar, B-On e RCAAP. We analyzed articles produced until 2015 inclusive. The methodological quality and its psychometric properties were accessed and summarized through two standardized criteria, following the COSMIN ideology.

Results: In this study they were included 6 articles. Disabilities of the Arm Shoulder and Hand and Neck and Upper Limb Index show good psychometric properties and an average quality; both Shoulder Pain And Disability Index and Weelchair User's Shoulder Pain Index exhibit good psychometric properties but poor methodological quality; both Shoulder Rating Questionnaire and Upper Extremity Functional Index relevant psychometric properties have not been evaluated, however the analyzed ones have good measurement properties and poor methodology.

Discussion and conclusions: Because of flaws in the methodology of the included studies, it is not possible to infer the most appropriate questionnaire in clinical practice. More studies are needed for the validation of patient reported outcome measures with better methodological quality.

Keywords: Shoulder complex, psychometric properties, patient reported outcome measures, COSMIN.

INTRODUÇÃO

Em Portugal, no período de um ano a prevalência total de dor no ombro varia de 14% para 21%. Dentro da população que sofre de dor músculo-esquelética, 18% dos pagamentos de seguro de invalidez remete para utentes com distúrbios cervicais e do ombro.[1] O tratamento de lesões referentes ao ombro é bem-sucedido se, para além de melhorar os problemas estruturais do indivíduo, recuperar também as suas atividades e participação na sociedade.[2]

Existem vários questionários de autorresposta para avaliar de utentes com disfunções no Complexo Articular do Ombro (CAO), porém a maioria foi desenvolvida na língua inglesa.[3–5] Contudo, apesar dos já validados para a população portuguesa e da importância destes na avaliação das disfunções do CAO, não foram encontradas revisões que efetuem a análise das propriedades psicométricas dos instrumentos validados para Portugal.[3–6] Este tipo de estudo demonstra ser essencial pois permite uma visão clara e compreensiva das propriedades de medida.[7] A seleção do instrumento adequado deve-se basear em grande parte na força das suas medidas psicométricas.[8] Assim levanta-se a questão: Qual/ais os instrumentos de medida de autorresposta relacionados com as disfunções do CAO que apresenta/m melhor/es propriedades psicométricas?

Relativamente às propriedades psicométricas, existe uma falta de consenso no que diz respeito à terminologia e definições. Ao longo desta revisão, será utilizada a nomenclatura proposta por Mokkink et al..[9] Para além disso, atualmente o grupo *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN) tem-se demonstrado relevante nas pesquisas da área da saúde para a definição das propriedades de medida e as COSMIN *checklists* têm contribuído para a revisão de escalas de avaliação relacionadas com a saúde.[10]

O presente estudo tem como objetivos: identificar estudos que se propõem investigar as propriedades psicométricas de instrumentos de autorresposta acerca de disfunções no CAO; avaliar a qualidade metodológica dos estudos; avaliar as propriedades psicométricas de cada instrumento; providenciar aos profissionais de saúde uma síntese da evidência acerca das propriedades psicométricas de cada instrumento; concluir acerca da qualidade dos instrumentos traduzidos e validados para a população portuguesa.

MATERIAL E MÉTODOS

TIPO DE ESTUDO

Este estudo consiste numa revisão sistemática e é classificado como qualitativo.[11] Para tal foi seguida a metodologia do estudo realizado por Huang et al., as recomendações do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) no desenvolvimento da revisão sistemática, e mais especificamente o protocolo desenvolvido pela Terwee, seguindo a ideologia da COSMIN.[7,12]

ESTRATÉGIA DE BUSCA

A revisão da literatura foi realizada em inglês/português, nas seguintes bases de dados: PubMed; PEDro; Google Académico; B-On e Repositório Científico de Acesso Aberto de Portugal (RCAAP). Com as palavras-chave: "*Assessment instrument*"; "*Assessment scale*"; "*measurement scale*"; "*Outcome measure*"; "PROM"; "*Evaluation*"; "*Psychometric properties*"; "*Reliability*"; "*Validity*"; "*Responsiveness*"; "*Shoulder injury*"; "*Upper limb*"; "*Portuguese*"; "Portugal"; "Instrumento de avaliação"; "Instrumento de medida"; "Avaliação"; "Instrumentos de autorresposta"; "Propriedades psicométricas"; "Validade"; "Confiabilidade"; "Reprodutibilidade"; "Capacidade de Resposta"; "Ombro"; "Membro superior". Foram incluídos artigos/trabalhos académicos realizados até ao ano de 2015, inclusive.

CRITÉRIOS DE SELEÇÃO

Os critérios de inclusão foram: serem artigos/trabalhos académicos sobre propriedades psicométricas de questionários de autorresposta relacionados com disfunções no membro superior/ombro; apresentarem pelo menos uma das características psicométricas; a escala/questionário do artigo/trabalho académico estar traduzida para português-europeu e adaptado para a população portuguesa. E foram excluídos: artigos em que a população fosse predominantemente constituída por crianças (<18 anos); não ter acesso ao artigo/trabalho académico completo.

INSTRUMENTOS

Avaliação da Qualidade Metodológica

Para atenuar diferenças na avaliação da qualidade metodológica dos estudos sobre propriedades de medida foi desenvolvida a COSMIN *checklist.*[13] Posteriormente foi desenvolvido um sistema de pontuação para calcular a qualidade metodológica por cada propriedade de medida presente na COSMIN *checklist* – a COSMIN *checklist* modificada (com escala de 4 pontos).[14,15]

Assim, tal como recomendado por Terwee, nesta revisão sistemática foi utilizada a COSMIN *checklist* modificada como instrumento de avaliação da qualidade metodológica dos estudos incluídos. Esta avaliação foi realizada por duas revisoras independentes para cada um dos artigos. Nos casos de discordância houve tentativa de consenso. Se este não se verificasse uma terceira pessoa era consultada.[7]

Avaliação das Propriedades Psicométricas

Como forma de avaliar as propriedades psicométricas da literatura selecionada, foi utilizada uma escala de classificação proposta inicialmente por Terwee et al. e modificada por Schellingerhout et al..[16,17] Esta apresenta como vantagem a utilização da nomenclatura definida pela COSMIN.[9] Os critérios encontram-se descritos na Tabela 1.

Métodos de síntese

Para sintetizar todos os dados recolhidos e chegar a uma categorização geral da evidência, foi utilizado um método proposto por Terwee.[7] Neste método, a síntese dos estudos é realizada combinando o número e a qualidade metodológica dos estudos com a consistência da classificação da evidência psicométrica baseado nos níveis de evidência propostos pelo *Cochrane Back Review Group* (Tabela 2).[18]

RESULTADOS

Neste estudo foram incluídos 6 artigos sobre as propriedades psicométricas de 6 instrumentos de autorresposta de avaliação de disfunções no CAO (Imagem 1). Para cada um dos instrumentos foi encontrado apenas um estudo de validação.

A caracterização dos estudos incluídos bem como a descrição dos questionários encontram-se resumidos na Tabela 3 e na Tabela 4, respetivamente. Após consenso na avaliação

metodológica resultaram os dados apresentados na Tabela 5. Nesta encontra-se ainda a avaliação de cada propriedade psicométrica segundo os critérios de qualidade descritos na Tabela 1. Os resultados obtidos foram sintetizados segundo os níveis de evidência (Tabela 2) e encontram-se resumidos na Tabela 6.

Disabilities of the Arm Shoulder and Hand (DASH)[19]

A consistência interna do DASH apresenta um alfa de *Cronbach* de 0.95 e, segundo a classificação COSMIN, uma qualidade metodológica fraca. Desta forma, para a propriedade consistência interna a evidência é desconhecida. Relativamente à confiabilidade esta foi calculada através da correlação de *Pearson*, obtendo um valor de 0.886, apresentando uma qualidade metodológica média. Assim, o nível de evidência para esta propriedade é limitado positivo. No teste de hipóteses, os autores analisaram as correlações existentes entre o DASH, a severidade da dor e grau de incapacidade, verificando o que inicialmente tinham conjeturado, obtendo valores de correlação de 0.49 para a dor e de 0.54 para a incapacidade. O teste de hipóteses apresenta uma qualidade metodológica fraca, obtendo assim um nível de evidência desconhecido. Relativamente à validade transcultural foram realizados todos os passos de tradução, no entanto, apresenta uma qualidade metodológica fraca. Para a validade de critério os autores correlacionaram as pontuações obtidas no DASH com as pontuações obtidas no *Short-form Health Survey 36-Item* (SF-36), obtendo uma correlação negativa e significativa, não apresentando o valor desta correlação. Contudo a qualidade metodológica para a validade de critério é classificada como boa.

Neck and Upper Limb Index (NULI-20)[20]

A consistência interna tem um valor de alfa de *Cronbach* de 0.92 para o questionário global. Segundo a classificação da COSMIN, a consistência interna deste estudo é considerada de fraca qualidade metodológica, tendo um nível de evidência desconhecido. No estudo foi encontrado um valor de confiabilidade de Coeficiente Correlação Intraclasse (CCI) igual a 0.83, esta propriedade é considerada de média qualidade metodológica e um nível de evidência limitado positivo. Quanto à validade de conteúdo, os indivíduos entrevistados no estudo consideraram o NULI-20 claro, compreensível e adequado à sua condição, obtendo assim uma qualidade metodológica excelente e um nível de evidência forte positivo. O teste de hipóteses foi avaliado através do coeficiente de correlação de *Spearman*, com o valor de 0.612, correspondendo a uma qualidade metodológica média e a um nível de evidência limitado positivo. Relativamente à validade transcultural foram realizados todos os passos de

tradução, sendo avaliada com qualidade metodológica fraca. A validade de critério foi avaliada através do coeficiente de correlação de *Spearman* este varia entre -0.340 e -0.688, avaliada com boa qualidade metodológica e um nível de evidência moderado negativo. A capacidade de resposta foi avaliada através da medida estatística *standardized effect size* (ES) apresentando-se com ES =0.95, correspondendo a uma qualidade metodológica fraca e a um nível de evidência desconhecido.

Shoulder Pain And Disability Index (SPADI)[21]

A consistência interna, calculada pelo alfa de *Cronbach*, apresenta um valor de 0.75 para a dimensão dor e 0.84 para atividade funcional. Quanto à confiabilidade foi aplicada a correlação de *Pearson*, sendo obtidos valores de 0.898 para a dimensão da dor e de 0.861 para a dimensão atividade funcional. Quanto à validade de conteúdo, os indivíduos entrevistados no estudo foram questionados apenas acerca da compreensão da primeira pergunta de cada dimensão. O teste de hipóteses foi calculado pela análise das correlações das diferentes dimensões e a idade, severidade da dor e tempo de doença. Relativamente à idade não se verificaram correlações significativas; para a severidade da dor verificaram-se valores de 0.730 para a dor e de 0.490 para a atividade funcional; para o tempo de doença verificaram-se valores de 0.494 para a dor e de 0.532 para a atividade funcional. Quanto à validade de critério, esta foi avaliada através da correlação das pontuações obtidas no SPADI com as do SF-36. Dentro da dimensão dor no SPADI foram obtidos valores de correlação que variam de -0.154 a -0.655; dentro da dimensão da atividade funcional estes variam entre –0.087 e -0.801. Relativamente à capacidade de resposta, esta foi dada pela correlação dos valores médios das diferenças verificadas nos dois momentos de avaliação entre a pontuação do SPADI, do SF-36 e as amplitudes de movimento, não se tendo verificado qualquer tipo de correlações. Assim todas as propriedades psicométricas apresentam qualidade metodológica fraca de acordo com a classificação COSMIN, resultando num nível de evidência desconhecido.

Shoulder Rating Questionnaire (SRQ-PT)[22]

Para a consistência interna foi calculado o alfa de *Cronbach*, que obteve um valor de 0.91 no questionário total. Foi classificada segundo a COSMIN com qualidade metodológica fraca e apresentando um nível de evidência desconhecido. Quanto à confiabilidade foi utilizado o coeficiente de correlação de *Spearman* obtendo um valor de 0.90 para o questionário total. Esta propriedade apresenta fraca qualidade metodológica e um nível de evidência

desconhecido. A equivalência semântica e de conteúdo entre as versões, foram efetuadas através de todos os passos de tradução. A validade de conteúdo é classificada como sendo de excelente qualidade metodológica e com um nível de evidência forte positivo. A validade transcultural é de fraca qualidade metodológica.

Upper Extremity Functional Index (UEFI)[23]

Os resultados dos testes de consistência interna apresentam um valor de alfa de *Cronbach* de 0.93. Segundo a COSMIN esta propriedade é classificada como fraca na sua qualidade metodológica, e com um nível de evidência desconhecido. A confiabilidade foi calculada através da correlação de *Pearson* apresentando valores que variam entre 0.61 e 0.972, sendo considerada de fraca qualidade metodológica e com um nível de evidência desconhecido. Relativamente à validade transcultural foi efetuado todo o processo de tradução, contudo apresenta qualidade metodológica fraca. Quanto à validade de conteúdo, resultou o consenso de que o questionário é útil e adequado à população a que se dirige. Desta forma foi classificado com excelente qualidade metodológica e de nível de evidência forte positivo.

Weelchair User's Shoulder Pain Index (WUSPI)[24]

A consistência interna apresenta um valor de alfa de *Cronbach* de 0.907. A confiabilidade apresenta um valor de correlação de *Pearson* de 0.998. Relativamente à validade de conteúdo, os indivíduos entrevistados no estudo consideraram o WUSPI claro, compreensível e adequado à sua condição. Este apresenta qualidade metodológica excelente e um nível de evidência forte positivo. No teste de hipóteses, foram analisadas as correlações existentes entre o WUSPI e os dados sociodemográficos, os dados relativos à utilização de cadeira de rodas e os dados relativos à sintomatologia dolorosa no ombro sendo que apenas se verificaram correlações entre o WUSPI e alguns aspetos relativos à sintomatologia. A validade transcultural foi realizada através de todos os passos de tradução. Em relação à validade de critério, correlacionaram, através da correlação de *Pearson,* o WUSPI com a escala de severidade da dor no ombro, as amplitudes de movimento ativo do ombro e o estado de saúde utilizando o SF-36. Relativamente à severidade da dor esta apresenta uma correlação com o WUSPI que varia de 0.929 a 0.825. No que respeita às correlações com as amplitudes de movimento ativo do ombro estas apresentam valores entre -0.416 e -0.912. Em relação ao SF-36 não foram verificadas correlações significativas com o WUSPI. A capacidade de resposta foi avaliada através da análise das pontuações obtidas em três momentos. Os autores analisaram as correlações entre as médias das alterações encontradas

no WUSPI com as médias das alterações encontradas na severidade da dor, nas amplitudes de movimento ativo e no estado de saúde. Assim as correlações existentes para a severidade da dor apresentam valores de 0.474 entre t_0-t_1 e de 0.511 entre t_1-t_2. As correlações existentes para as amplitudes de movimento ativo variam entre 0.380 e 0.759 para t_0-t_1 e entre 0.165 e 0.733 para t_1-t_2. Para as médias das alterações verificadas no estado de saúde avaliado através do SF-36 não se verificaram correlações significativas com as médias das alterações do WUSPI. Todas as propriedades psicométricas, à exceção da validade de conteúdo, são consideradas, segundo a classificação COSMIN, como fracas na sua qualidade metodológica e o nível de evidência é desconhecido.

DISCUSSÃO

O DASH tem sido utilizado na avaliação do membro superior, no entanto este é frequentemente utilizado em condições relacionadas com o ombro. Para comparar dois tipos de intervenção na capsulite adesiva do ombro, foi realizado um estudo, onde a forma de avaliação dos resultados foi o DASH.[25] Num estudo realizado a indivíduos com dor no ombro não traumática, o DASH foi utilizado para verificar a efetividade dum tratamento no alívio da dor.[26] Noutro estudo sobre dor no ombro relacionada com capsulite adesiva, o DASH foi também utilizado como forma de avaliar a eficácia dum medicamento nesta condição.[27]

Relativamente aos resultados do nosso estudo, na metodologia do artigo referente ao DASH, a consistência interna caracterizou-se como fraca, uma vez que não foi calculado para cada dimensão o valor de alfa de *Cronbach*. Já a confiabilidade classificou-se como média, considerando que não foram descritas as condições de teste nem o estado dos pacientes ao longo do tempo de aplicação. Em relação à validade de conteúdo, esta não foi avaliada no nosso estudo uma vez que os dados apresentados no artigo eram subjetivos. Quanto ao teste de hipóteses, este foi considerado fraco dado que não foi fornecida uma descrição adequada do instrumento comparador. No que concerne à validade de critério, esta foi classificada como boa, sendo que não foi fornecida evidência que o critério utilizado poderia ser considerado um *gold standard*. Apesar das falhas metodológicas, a avaliação das propriedades psicométricas apresentou uma pontuação positiva para todas as componentes.

O NULI-20 avalia o estado funcional de trabalhadores com lesões músculo-esqueléticas (cervical e membro superior). Contudo, foi realizado um estudo que se propôs avaliar a efetividade de um protocolo de exercícios de relaxamento muscular em músicos com

alterações ao nível do ombro. Desta forma, o NULI-20 foi aplicado com vista à perceção da redução da dor e desconforto nos músculos do CAO, nomeadamente ao nível dos trapézios e deltoïdes.[28]

Relativamente aos resultados do nosso estudo, aquando da avaliação da metodologia do artigo referente ao NULI-20, a consistência interna evidenciou-se como fraca uma vez que a análise fatorial não foi realizada e não referem outro estudo, bem como a ausência do cálculo do ajuste estatístico a nível global. Em relação à confiabilidade, esta foi classificada como média devido ao tamanho da amostra. O teste de hipóteses foi classificado como médio, dado que apenas foram dadas algumas informações sobre as propriedades de medida do instrumento comparador, mas não específico para esta população. A validade de critério foi qualificada como boa devido ao tamanho da amostra em estudo. A capacidade de resposta foi avaliada como fraca uma vez que o intervalo de tempo não foi descrito adequadamente. Todas as propriedades psicométricas foram classificadas como positivas exceto a validade de critério que foi negativa uma vez que a correlação com o *gold standard* (*Short-form Health Survey 12-Item* (SF-12)) apresentava um valor inferior a 0.7.

O SPADI tem sido bastante utilizado na avaliação da funcionalidade e dor no ombro. Num estudo para determinar a eficácia da proloterapia em utentes com doença crónica não traumática da coifa dos rotadores, foi utilizado o SPADI como instrumento de avaliação.[29] Num estudo comparativo entre duas intervenções, foi utilizado o SPADI por forma a avaliar as alterações na funcionalidade e dor no ombro em utentes com tendinopatia da coifa dos rotadores.[30] Noutro estudo foi utilizado o SPADI para acompanhar as alterações na dor e funcionalidade de utentes com dor no ombro após terem sido intervencionados em cuidados primários.[31]

No nosso estudo, o SPADI exibiu uma consistência interna de fraca qualidade metodológica devido à ausência da análise fatorial sem referência a outro estudo e ao facto de o ajuste estatístico a nível global não ter sido calculado. A validade de conteúdo evidenciou-se fraca, dado que não se verificou se os itens em conjunto refletiam o constructo a ser medido e se eram relevantes para a população. O teste de hipóteses classificou-se como fraco por não ter sido apresentada nenhuma informação acerca das propriedades do instrumento comparador. Esta classificação fraca do instrumento, em algumas das propriedades, deve-se ao tamanho da amostra, sendo que participaram apenas 29 pessoas. Contudo, segundo os critérios da COSMIN, a classificação subiria para média na avaliação da metodologia da confiabilidade, validade de critério e capacidade de resposta se a amostra fosse superior a 30. Todas as

propriedades foram avaliadas com classificação positiva exceto a validade de critério e a capacidade de resposta. Na validade de critério a maioria das correlações existentes são inferiores a 0.7 e na capacidade de resposta pela inexistência de correlações entre as pontuações do SPADI e as pontuações dos instrumentos comparadores.

O SRQ-PT avalia a sintomatologia e a funcionalidade do ombro. Foi realizado um estudo para comparar a eficácia de quatro questionários do ombro na perceção da dor em cuidados de saúde primários, de entre os quais o SRQ-PT foi um dos questionários que apresentou maior capacidade de resposta.[32] Outro estudo recorreu à aplicação do SRQ-PT a fim de avaliar a eficácia dum programa de exercícios terapêuticos destinados a reduzir a dor e melhorar a função no ombro.[33]

No momento de avaliação da metodologia do artigo referente ao SRQ-PT, a consistência interna revelou ser fraca dado que a análise fatorial não foi realizada, não havendo referência a outro estudo e o ajuste estatístico a nível global também não foi calculado. A confiabilidade foi considerada fraca dado que as condições de teste não foram similares. Relativamente à avaliação das propriedades psicométricas segundo os critérios de qualidade, a confiabilidade é indeterminada pois o valor de alfa de *Cronbach* não foi calculado, sendo calculada utilizando o coeficiente de correlação de *Spearman*, para o qual o valor foi de 0.90.

Embora o UEFI avalie o membro superior, este tem sido utilizado nalguns estudos especificamente na avaliação da funcionalidade do ombro. O UEFI foi utilizado para avaliar a funcionalidade do ombro num estudo em que foi efetuada uma comparação entre duas intervenções.[34] Noutro estudo, no qual foi comparada a capacidade de resposta de diversos instrumentos de autorresposta em utentes com patologias da coifa dos rotadores, verificaram que o UEFI é o que apresenta melhor capacidade de resposta.[35]

No nosso estudo na avaliação da metodologia do artigo referente ao UEFI, a consistência interna foi fraca devido ao tamanho da amostra, a análise fatorial não foi realizada, não mencionando outro estudo e não foi calculado um ajuste estatístico global. A confiabilidade é fraca, devido ao tamanho da amostra. Na avaliação das propriedades psicométricas do UEFI, estas foram consideradas todas como positivas.

O WUSPI pretende medir a intensidade da dor no ombro em utilizadores de cadeira de rodas durante a execução de atividades da vida diária. Num estudo, com o objetivo de identificar a relação de dor no ombro com qualidade de vida, atividade física e atividades sociais em

pessoas com paraplegia, utilizaram como métodos de avaliação diferentes instrumentos um dos quais o WUSPI.[36] Num estudo para verificar a efetividade de uma intervenção, foi utilizado o WUSPI na avaliação da dor no ombro.[37] Noutro estudo para testar a eficácia do exercício na dor no ombro foram utilizados como forma de medição o WUSPI, o DASH e o SRQ.[38]

Na avaliação da metodologia do artigo referente ao WUSPI, a consistência interna foi fraca pelo facto da análise fatorial não ter sido realizada, não referindo outro estudo e pelo ajuste estatístico global não calculado. Neste estudo, à semelhança do SPADI, teria sido classificado com uma melhor qualidade metodológica (média) nas propriedades de confiabilidade, teste de hipóteses, validade de critério e capacidade de resposta se a amostra utilizada tivesse incluído mais indivíduos. As propriedades psicométricas do WUSPI foram todas consideradas como positivas à exceção do teste de hipóteses e da validade de critério. O primeiro foi considerado como negativo dada a existência de correlações não significativas em alguns parâmetros comparadores, não confirmando, pelo menos, 75% das hipóteses formuladas inicialmente. A validade de critério é negativa porque não existem correlações significativas entre os valores obtidos entre o WUSPI e o SF-36.

De um modo global, o tamanho da amostra incluída revelou-se reduzida, diminuindo a qualidade metodológica dos estudos, sendo esta uma característica transversal a todos os instrumentos. Além disso, verificou-se que nenhum dos estudos analisou o erro de medida nem a validade estrutural. Outro aspeto relaciona-se com as falhas metodológicas dos estudos dado que, por vezes, não foram realizadas as análises estatísticas adequadas ou estas não se encontravam descritas.

OUTRAS CONSIDERAÇÕES

Numa revisão sistemática realizada por Bot et al., que reuniu 16 escalas relativas à avaliação da incapacidade e função física do ombro, a maioria dos estudos incluídos referem-se ao DASH e ao SPADI. Segundo os resultados desse estudo, o DASH apresenta melhor classificação das propriedades de medida.[4] Numa revisão sistemática de Roy et al., após a análise das propriedades de medida de 4 instrumentos de incapacidade do ombro, entre os quais foram incluídos o DASH e o SPADI, os resultados evidenciaram que todos os questionários são aceitáveis para uso clínico.[5] Ao contrário do nosso estudo, nestas revisões sistemáticas, os estudos incluídos demonstraram processos de validação mais rigorosos, sendo possível aferir uma classificação a cada um dos questionários estudados. Contudo, à

semelhança do nosso estudo, numa revisão sistemática realizada por Puga et al. também não foi possível concluir acerca da qualidade dos instrumentos uma vez que as propriedades psicométricas não foram testadas da forma mais correta.[3]

LIMITAÇÕES E PONTOS FORTES

Quanto às limitações do presente estudo destaca-se o número reduzido de artigos. Para além disso poucos são os artigos que foram submetidos a revisão e publicados. Salienta-se ainda a falta de um *expert* no grupo.

Quanto aos pontos fortes temos que o nosso estudo é inédito em Portugal, fornecendo um instrumento útil aos profissionais de saúde. O uso da nova nomenclatura baseada em critérios internacionais, bem como o uso de critérios de qualidade padronizados utilizados na elaboração de revisões sistemáticas de propriedades psicométricas e a utilização da ferramenta PRISMA constitui outro ponto forte.

CONCLUSÕES

Embora a maioria dos instrumentos apresente boas características psicométricas, a metodologia dos estudos incluídos revela algumas lacunas, classificando-se na maioria das vezes como fraca. Desta forma, não é viável considerar os resultados dos estudos como totalmente fidedignos, não sendo possível inferir qual o questionário mais apropriado à prática clínica dos profissionais de saúde. Aquando da escolha de um instrumento de autorresposta, os profissionais de saúde deverão estar atentos à qualidade metodológica das escalas e não apenas à qualidade das propriedades psicométricas. São necessários mais estudos de tradução e validação de instrumentos de autorresposta em Portugal com qualidade metodológica mais elevada.

RECOMENDAÇÕES PARA O FUTURO

Dado que os instrumentos já traduzidos e validados para a população portuguesa apresentam metodologia fraca, é recomendada a realização de novos estudos de validação para os mesmos. Recomenda-se a realização de estudos semelhantes para outros constructos.

BIBLIOGRAFIA

1. APED. *Dor No Ombro.*; 2010.

2. Paternostro-Sluga T, Zöch C. Konservative Therapie und Rehabilitation von

Schultererkrankungen. *Schulterdiagnostik*. 2004:597-603.

3. Puga VO, Lopes AD, Costa LO. Assessment of cross-cultural adaptations and measurement properties of self-report outcome measures relevant to shoulder disability in Portuguese: a systematic review. *Rev Bras Fisioter*. 2012;16(2):85-93.

4. Bot SDM, Terwee CB, van der Windt DAWM, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Ann Rheum Dis*. 2004;63(4):335-341.

5. Roy J, MacDermid J, Woodhouse L. Measuring Shoulder Function : A Systematic Review of Four Questionnaires. *Arthritis Rheum (Arthritis Care Res*. 2009;61(5):623-632.

6. Silva M. Resultados de Medida. *Essfisionline*. 2006;2:59-72.

7. Terwee C. Protocol for systematic reviews of measurement properties. *Meas Med*. 2011.

8. Kyte DG, Calvert M, van der Wees PJ, ten Hove R, Tolan S, Hill JC. An introduction to patient-reported outcome measures (PROMs) in physiotherapy. *Physiother (United Kingdom)*. 2015;101(2):119-125.

9. Mokkink L, Terwee C, Patrick D, 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*. 2010;63(7):737-745.

10. Polit DF. Assessing measurement in health: Beyond reliability and validity. *Int J Nurs Stud*. 2015;52(11):1746-1753.

11. Evans D, Pearson A. Systematic reviews of qualitative research. *Clin Eff Nurs*. 2001;5(3):111-119.

12. Huang H, Grant JA, Miller BS, Mirza FM, Gagnier JJ. A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease. *Am J Sports Med*. 2015;43(10):2572-2582.

13. Mokkink LB, Terwee C, Patrick D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status

measurement instruments: An international Delphi study. *Qual Life Res.* 2010;19(4):539-549.

14. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, Vet HCW de. Rating the methodological quality in systematic reviews of studies on measurement properties : a scoring system for the COSMIN checklist. *Qual Life Res.* 2012:651-657.

15. Terwee C. COSMIN checklist with 4-point scale. COSMIN.

16. Terwee CB, Bot SDM, Boer MR De, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34-42.

17. 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. *Qual Life Res.* 2012;21(4):659-670.

18. van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the cochrane collaboration Back Review Group. *Spine (Phila Pa 1976).* 2003;28(12):1290-1299.

19. Santos J, Gonçalves R. Adaptação e validação cultural da versão portuguesa do Disabilities of the Arm Shoulder and Hand–DASH. *Rev Port Ortop e Traumatol.* 2006;14:29-46.

20. Matias S et al. *Tradução E Adaptação Cultural Do Neck and Upper Limb Index Para a Língua Portuguesa.*; 2010.

21. Duarte A. *Validação Intercultural Do Shoulder Pain and Disability Index- SPADI.*; 2002.

22. Guerreiro JA, Proença I, Moura N, Cartucho A. Adaptação transcultural do Shoulder Rating Questionnaires para a Língua portugues (SRQ-PT): Tradução; Validação; Análise da consistência interna e replicabilidade. *ifisionline.* 2011;1(2):5-18.

23. Melo F. *Upper Extremity Functional Index- Adaptação Cultural E Linguística.*; 2002.

24. Clara F. *Validação Intercultural Do Wheelchair User's Shoulder Pain Index.*; 2001.

25. Ibrahim M, Donatelli R, Hellman M, Echternach J. Efficacy of a static progressive stretch device as an adjunct to physical therapy in treating adhesive capsulitis of the shoulder: A prospective, randomised study. *Physiother (United Kingdom)*. 2013;100(3):228-234.

26. Bron C, de Gast A, Dommerholt J, Stegenga B, Wensing M, Oostendorp R a B. Treatment of myofascial trigger points in patients with chronic shoulder pain: a randomized, controlled trial. *BMC Med*. 2011;9(1):8-22.

27. Buchbinder R, Hoving JL, Green S, Hall S, Forbes A, Nash P. Short course prednisolone for adhesive capsulitis (frozen shoulder or stiff painful shoulder): a randomised, double blind, placebo controlled trial. *Ann Rheum Dis*. 2004;63(11):1460-1469.

28. Rodrigues A, Santana M, Pinheira V. *Avaliação Da Efetividade de Um Protocolo de Exercícios de Relaxamento Muscular Em Músicos Com Alterações Músculoesqueléticas.*; 2014.

29. Lee D-H, Kwack K-S, Rah Woo U, Yoon S-H. Prolotherapy for Refractory Rotator Cuff Disease: Retrospective Case-Control Study of 1-Year Follow-Up. *Arch Phys Med Rehabil*. 2015;96(11):2027-2032.

30. Littlewood C, Malliaras P, Mawson S, May S, Walters SJ. Self-managed loaded exercise versus usual physiotherapy treatment for rotator cuff tendinopathy: A pilot randomised controlled trial. *Physiother (United Kingdom)*. 2014;100(1):54-60.

31. Laslett M, Steele M, Hing W, McNair P, Cadogan A. Shoulder pain patients in primary care - Part 1: Clinical outcomes over 12 months following standardized diagnostic workup, corticosteroid injections, and community-based care. *J Rehabil Med*. 2014;46(9):898-907.

32. Paul A, Lewis M, Shadforth MF, Croft PR, Van Der Windt D a WM, Hay EM. A comparison of four shoulder-specific questionnaires in primary care. *Ann Rheum Dis*. 2004;63(10):1293-1299.

33. Ludewig P, Borstad J. Effects of a home exercise programme on shoulder pain and functional status in construction workers. *Occup Environ Med*. 2003;60(11):841-849.

34. Kolmus AM, Holland AE, Byrne MJ, Cleland HJ. The effects of splinting on shoulder function in adult burns. *Burns*. 2012;38(5):638-644.

35. Razmjou H, Bean A, van Osnabrugge V, MacDermid JC, Holtby R. Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. *Bmc Musculoskelet Disord*. 2006;7(1):26-33.

36. Gutierrez DD, Thompson L, Kemp B, Mulroy SJ. The relationship of shoulder pain intensity to quality of life, physical activity, and community participation in persons with paraplegia. *J Spinal Cord Med*. 2007;30(3):251-255.

37. Nawoczenski D a, Ritter-Soronen JM, Wilson CM, Howe B a, Ludewig PM. Clinical trial of exercise for shoulder pain in chronic spinal injury. *Phys Ther*. 2006;86(12):1604-1618.

38. Straaten M, Cloud B, Morrow M, Ludewig P, Zhao K. Effectiveness of Home Exercise on Pain, Function, and Strength of Manual Wheelchair Users With Spinal Cord Injury: A High-Dose Shoulder Program With Telerehabilitation. *Arch Phys Med Rehabil*. 2014;95(10):1810-1817.