AVANCES EN CIRUGÍA DE LA MANO Y USO DE INSTRUMENTOS PRO PARA MEDIR RESULTADOS

Roberto Sánchez Rosales

Unidad de Cirugía de La Mano y Microcirugía, GECOT. La Laguna, Tenerife. España. María del Cristo Osunna 20 , La Laguna 38204. +34637411872 titorosaleselefonica.net

RESUMEN

La Clasificación Mundial del Funcionamiento, la Discapacidad y la Salud (ICF) fue desarrollada por la Organización Mundial de la Salud para su aplicación en diversos aspectos de la salud. Las medidas de la salud pueden clasificarse según la ICF en: función/ estructura del cuerpo, actividad y participación. Tradicionalmente, las medidas utilizadas en la cirugía de la mano habían sido principalmente en la función corporal y estructura, como radiografías, rango de movimiento , 2 puntos de discriminación sensitiva, etc . Recientemente , el uso de instrumentos de salud basados en la opinión del paciente (PRO : "Patient Reported Outcomes") ha sido introducido en cirugía de la mano para valorar resultados siendo considerados como medidas de la actividad y participación según la ICF . El propósito de este trabajo de revisión fue el presentar los conceptos básicos y requerimientos metodológicos en el uso de instrumentos PRO que ha servido para mejorar la calidad de la investigación clínica en cirugía de la mano.

ABSTRACT

The International Classification of Functioning, Disability and Health (ICF) was developed by the World Health Organization for application to various aspects of health. Health measures can be classified according to the ICF into body function and structure, activity, and participation. Traditionally, the measures used in hand surgery have been mainly on body function and structure level such as radiographs, strength measures (grip, key pinch , pulp to pulp pinch) , range of motion (ROM), two points discrimination (2ppd), etc. The use of health instruments or questionnaires based on the patients 'opinion, called Patients Reported Outcomes (PRO), have been introduced in hand surgery. They can be considered as measures on activity and participation based on ICF classification.The purpose of this review paper was to present the basic concepts and methodological requirements in the use of PRO instruments which helped to improve the quality of clinical research in hand surgery

INTRODUCCIÓN

Desde la primera sutura microscópica de arterias digitales o de pequeño calibre por Harold Kleinert (1), la cirugía de la mano ha alcanzado límites impensables. Hoy se realizan de manera sistemática: reimplantes, colgajos libres vascularizados, injertos nerviosos, transferencias de dedo de pie a mano, e incluso trasplantes de mano de cadáver. Este desarrollo técnico no se había acompañado de una mejora en la calidad de los trabajos de investigación clínica publicados en esta área de la cirugía. Así, se seguía confundiendo incidencia con prevalencia cuando se hablaba de la ocurrencia de una enfermedad o patología crónica en la mano (2-5); se utilizaba la palabra "retrospectivo" en el título de los estudio (6), cuando tanto un diseño clínico descriptivo como una serie de casos o un estudio de corte transverso, un estudio de casos control y un diseño cohortes histórico son todos retrospectivos, pero con niveles de evidencia

científica diferentes (7) ; el número de ensayos clínicos y estudios observacionales publicados en las principales revistas científicas relacionadas con la mano y miembros superiores (MMSS), era muy bajo y de muy poca calidad (7-9), lo que repercutía a la hora de la revisión sistemática de la evidencia científica, haciendo casi imposible la realización de meta-análisis y el establecimiento de guías clínicas y protocolos de actuación ante patologías importantes en la cirugía de la mano, y repercutiendo por ende en el nivel de evidencia de los trabajos de investigación clínica publicados en cirugía de la mano (10, 11).

En la evaluación de la efectividad de una intervención, un aspecto importante es el tipo de medidas de resultados (Outcomes) utilizadas y si cubren todos los aspectos importantes del efecto del tratamiento. La Clasificación Internacional de Funcionamiento, Discapacidad y Salud (ICF) fue desarrollada por la Organización Mundial de la Salud (12) para su aplicación a diversos aspectos de la salud. Las medidas de resultados se pueden clasificar según la ICF en función/estructura del cuerpo, actividad y participación. Tradicionalmente, las medidas utilizadas para evaluar los resultados en la cirugía de la mano han sido principalmente de función/estructura del cuerpo, como radiografías, medidas de fuerza y rango de movimiento (10). Recientemente, se ha introducido el uso de instrumentos salud basados en la opinión del paciente para medir resultados o Instrumentos PRO ("Patient-Reported Outcomes"), consideradas como medidas de la actividad y participación, lo que ha implicado una mejora en la valoración de los resultados así como en la calidad de los estudios clínicos en cirugía de la mano.

El propósito del siguiente trabajo de revisión ha sido presentar una actualización de los requerimientos metodológicos en el uso de instrumentos PRO aplicados en cirugía de la mano.

TIPOS DE INSTRUMENTOS PRO

Se pueden clasificar en dos grandes grupos en base al tipo de información recolectada: genéricos y específicos.

Los instrumentos genéricos, como el conocido "SF-36" o el "Euroqol 5D Index" son cuestionarios que tratan de medir todas las dimensiones importantes de la salud en relación a la calidad de vida, y pueden ser usados virtualmente en cualquier tipo de afección o patología independientemente de la alteración subyacente. Suelen tener menor respuesta o sensibilidad para detectar cambios de importancia clínica tras tratamiento, pero permiten la comparación entre diferentes enfermedades o trastornos lo cual puede ser de interés para los organizadores de la sanidad.

Los instrumentos específicos se centran en problemas asociados a una afección o patología específica, o a grupos o poblaciones de pacientes, o áreas de función. Presentan una mayor respuesta o sensibilidad para detectar cambios de importancia clínica tras tratamiento; pero no permiten comparación de resultados a través de diferentes enfermedades o afecciones. Dentro de este tipo de cuestionarios son de destacar por ejemplo: el cuestionario específico para el síndrome del túnel carpiano (Instrumento CTS) (Brigham and Women´s Hospital. Boston Carpal Tunnel Instrument) (13, 14) y su versión corta (CTS-6) (15-17) como representativos de instrumentos específicos de una enfermedad o afección patológica; el DASH (Disability of Arm, Shoulder and Hand), (Institute for Work & Health, AAOS, ASSH 1996) (13, 18, 19) que es un instrumento específico de un área de función, midiendo discapacidad de MMSS.; el PRWE (Patient Rated Wrist Evaluation) (20, 21) que mide discapacidad en relación a desórdenes de la muñeca,; y otros como el PEM (Patient Evaluation Measure), MHQ (Michigan Hand questtionnaire), etc.

PROPIEDADES DE LOS INSTRUMENTOS PRO

Los estándares basados en el consenso Delphi para la selección de instrumentos de medición de salud (COSMIN) (22, 23) se desarrollaron para evaluar la calidad de los estudios que involucran las propiedades de medición de un instrumento PRO. El COSMIN proporciona los estándares necesarios para el diseño y los análisis estadísticos recomendados. Se distinguen tres dominios a la

hora de analizar la calidad de un instrumento PRO : fiabilidad, validez y respuesta. Cada dominio contiene una o más propiedades de medición. Las propiedades de medición de las medidas PRO generalmente se evalúan en base a la teoría de prueba clásica (CTT), pero los métodos de la teoría de respuesta a ítems (IRT), como el análisis Rasch o el funcionamiento diferencial de ítems (DIF), se utilizan cada vez más para desarrollar y evaluar las propiedades psicométricas. de medidas PRO (24, 25). Como regla general, cualquier estudio que use una medida PRO debe incluir información sobre los siguientes problemas de diseño: ¿se informó el porcentaje de ítems perdidos o no contestados? ¿se describió el método de manejo de los ítems perdidos? ¿se describió la estimación del tamaño de la muestra? ¿se detectaron problemas con el diseño clínico?

FIABILIDAD

El dominio "fiabilidad" incluye: consistencia interna, fiabilidad test-re test y error de medición.

La consistencia interna representa el grado de interrelación entre los ítems que constituyen una escala. Se espera que los ítems estén altamente correlacionados. La consistencia interna se evalúa con el coeficiente alfa de Cronbach (⍰> 0.7 considerado como buena consistencia interna) (21, 26, 27). Un valor muy alto (> 0,90) podría sugerir la redundancia de ítems y que algunos de los ítems podrían eliminarse teóricamente sin sacrificar la consistencia interna. Un requisito importante para la interpretación del coeficiente alfa de Cronbach es que la escala debe ser unidimensional. La unidimensionalidad de la escala se puede analizar mediante el análisis factorial (FA).

La fiabilidad test – re test se evalúa comparando las respuestas cuando la medida PRO se administra dos veces a la misma población diana después de un período llamado "tiempo de lavado". Las dos administraciones deben ser independientes y la primera no debe influir en la segunda. El tiempo de lavado debe ser lo suficientemente largo como para evitar el "sesgo de recuerdo". No hay consenso sobre un intervalo específico para la fiabilidad test-retest. El COSMIN consideró apropiado un intervalo de 2 semanas para evaluar la confiabilidad de las medidas PRO (22, 23). Los pacientes deben estar estables durante el tiempo de lavado y las condiciones deben ser similares para la primera y segunda administración. La prueba estadística recomendada para la confiabilidad test-retest dependerá del tipo de opciones de respuesta del instrumento PRO. Para puntuaciones continuas se prefiere el coeficiente de correlación intraclase (ICC) (21, 27). Para cuestionarios con respuestas categóricas o nominales, se recomienda el coeficiente Kappa de Cohen (28, 29), y para escalas ordinales se recomienda el Kappa ponderado (29, 30).

Error de medición. La precisión de la medición estima el error alrededor de la puntuación observada en un instrumento PRO, ya sea en un punto "de" tiempo (precisión de corte transversal) o "en" el tiempo (precisión longitudinal). La precisión de corte transversal se analiza con el error estándar de la medición (SEM), basado en el coeficiente alfa de Cronbach (SEM = DS multiplicado por la raíz cuadrada de (1 – alfa Cronbach)). La evaluación de la precisión longitudinal requiere que el instrumento PRO se haya administrado dos veces y las administraciones hayan sido independientes y en condiciones similares. El error de medición longitudinal se analiza con el error estándar de la diferencia de medición (SEMdiff) (SEMdiff = DS * raíz cuadrada (1 – ICC) * raíz cuadrada de 2) y el cambio mínimo detectable (MDC), con un nivel de confianza del 90% (MDC90 = SEM diff * 1.65) o 95% (MDC95 = SEM *1.96) (19, 21)

Otro estadístico apropiado para evaluar el error de medición longitudinal basado en CCT es el límite de acuerdo de Bland y Altman (LoA) (31) basado en un simple gráfico en el que se enfrenta en el eje Y la diferencia entre ambas medidas para cada sujeto (di = Yi - Xi) y en el eje X la media (mi = (Xi + Yi) / 2)) de ambas medidas para cada individuo. Si asumimos la distribución de normalidad de las diferencias, se espera que el 95% de las diferencias se encuentre entre los límites del intervalo. LoA y el MDC, ambos están directamente relacionados con el SEMdiff (22, 23). Los cambios dentro del LoA o más pequeños que el MDC se consideran un error de medición y los cambios fuera del LoA o más grandes que

el MDC se consideran un cambio real en la puntuación. Un problema importante cuando comparamos MDC es que este índice de confiabilidad absoluta depende de varios factores, incluyendo la población de estudio, el tiempo de lavado, el tiempo de seguimiento cuando se realizó el análisis test-retest y la varianza de los datos (19, 21-23). Otro punto importante es que el MDC está relacionado con la confiabilidad y el error de medición, y es diferente de los conceptos de cambio clínico mínimo importante (MIC) o diferencia clínica mínima importante (MID) (32, 33) relacionado con la interpretabilidad del instrumento PRO.

VALIDEZ

El domino "validez" contiene tres propiedades de medición: validez de contenido, validez de constructo y validez de criterio.

La validez de contenido de un instrumento PRO es el grado en que su contenido es un reflejo adecuado del constructo a medir. Se recomienda que un panel de expertos juzgue la relevancia de todos los ítems relacionados con el constructo a medir. Debe incluir una evaluación de si todos los ítems son relevantes para las características de la población del estudio, como edad, sexo, entorno, etc. Un alto número de valores sin respuesta en un ítem específico podría sugerir que el ítem no es relevante para la población objetivo. Un alto número de pacientes con la puntuación más baja posible, o "efecto suelo" (floor effect), o con la puntuación más alta posible, o "efecto techo"(ceiling effect), puede indicar que se necesitan mejores ítems con mayor capacidad de diferenciación (22).

Basado en el COSMIN, la propiedad de ***medición "validez de constructo"*** incluye tres aspectos: validez estructural, prueba de hipótesis y validez transcultural

La validez estructural se define como el grado en que las puntuaciones de una medida PRO son un reflejo adecuado de la dimensión a medir. La prueba apropiada para evaluar la validez estructural es el análisis factorial (FA). El FA puede ser exploratorio o confirmatorio, ambos podrían ser útiles para la evaluación de la validez estructural de una medida PRO. El COSMIN prefiere FA confirmatorio (CFA) en el que las hipótesis sobre el concepto y las dimensiones se formulan a priori y el análisis evalúa si los datos se ajustan a una estructura factorial predefinida. Esto es importante cuando el investigador quiere justificar las diferentes dimensiones que mide un instrumento determinado. Por ejemplo, Rodrigues et al. (34) sugirieron que el DASH (Instrumento PRO que mide discapacidad de MMSS) no era unidimensional y cuestionaron la validez del DASH para la evaluación de resultados en la enfermedad de Dupuytren. Otra aplicación del FA es la reducción de ítem para crear una forma más corta de un instrumento PRO, como el CTS-6 del cuestionario de Boston (15) o el QuickDASH del DASH (18).

Test de hipótesis de constructo es el grado en que las puntaciones de un instrumento PRO son consistentes con hipótesis basadas en el supuesto de que el instrumento PRO mide válidamente el constructo a medir (17, 21, 22). Las pruebas de hipótesis pueden basarse en las correlaciones posibles entre del instrumento PRO en estudio, con las puntuaciones de otros instrumentos, o bien, en diferencias entre grupos relevantes (validez de constructo de grupos conocidos) (19). Los requisitos metodológicos importantes son: (1) se deben establecer las hipótesis antes de la recopilación de datos, (2) se debe definir la dirección esperada de las hipótesis y (3) se debe definir la magnitud absoluta o relativa esperada de la correlación. Por ejemplo, para la evaluación de validez de constructo del PRWE en fractura de radio distal, Rosales et al. plantearon la hipótesis de que el PRWE (discapacidad en relación a la muñeca) tendría una fuerte correlación positiva con el QuickDASH (discapacidad de MMSS) y una correlación negativa moderada con el índice EQ-5D (salud general y calidad de vida) al inicio del estudio y a las 8 semanas del tratamiento (21). Para evaluar las diferencias entre los grupos conocidos, el investigador debe definir la magnitud esperada de la diferencia en lugar del valor de p y las diferencias estadísticamente significativas, ya que estas últimas dependen más del tamaño muestral y del poder estadístico del análisis. Por ejemplo, en el análisis de validez de

constructo de prueba de hipótesis, MacDermid et al. (20) plantearon la hipótesis de que la puntuación del DASH (discapacidad de MMSS) sería más alto en el grupo de pacientes con osteoartritis de articulaciones múltiples. comparado con aquellos afectos de osteoartritis localizada de la mano.

La validez transcultural es el grado en que un instrumento PRO traducido o adaptado transculturalmente es un reflejo adecuado de la versión original del instrumento. Los requisitos metodológicos para la validez transcultural propuestos en el COSMIN son similares a los utilizados en el proyecto de Evaluación Internacional de la Calidad de Vida (IQOLA), desarrollado para obtener las diferentes versiones del SF-36 (35-37). Debe incluir información sobre el proceso de traducción hacia adelante y hacia atrás (al menos dos traductores bilingües que trabajan de forma independiente), cómo se resolvieron las diferencias entre las versiones originales y adaptadas, y una prueba previa que implica comprensión, relevancia cultural e interpretación de la traducción. El COSMIN recomendó CFA como prueba estadística para la evaluación de la validez intercultural. Cuando se utiliza IRT, se recomienda el análisis DIF (22, 38).

La validez de criterio es el grado en que las puntuaciones de un instrumento PRO son un reflejo adecuado de un "estándar de referencia". El panel COSMIN llegó a un consenso de que no existe un estándar de oro para los instrumentos PRO. Debido a que en el área de la salud no existen "medidas -criterios o patrones válidos y establecidos, la validez de criterio raramente es estudiada. Un error conceptual común es confundir la validez de criterio con la validez de constructo de prueba de hipótesis. Así, Alexander et al. (39) intentando analizar la validez de un nuevo instrumento de medición para medir las modernas actividades de la mano en relación a las nuevas tecnologías (ej. el uso del móvil, etc.), establecieron un estudio de validez de criterio correlacionando las mediciones de este nuevo instrumento con el DASH, usando éste último como medida de criterio de la función de la mano; cuándo los propios autores que desarrollaron el DASH expresaron en sus publicaciones originales que les fue imposible analizar la validez de criterio del DASH porque no existían medidas-criterio en la salud (19). La única excepción para la validez de criterio es cuando la versión corta de un instrumento PRO se compara con la versión larga original (22, 23). En ese caso, la versión larga original puede considerarse el estándar de referencia. Idealmente, la misma muestra de pacientes completa la versión larga y, después de un tiempo de lavado, la versión más corta y la validez del criterio se analizan con un coeficiente de correlación. Por ejemplo, para la versión más corta del Instrumento CTS (CTS-6), Atroshi et al. (15) utilizaron una muestra de 213 pacientes que completaron el QuickDASH y la escala de síntomas de CTS revisada, y 116 pacientes también completaron la escala de gravedad de síntomas de CTS original (intervalo medio de 11 días).

RESPUESTA

Respuesta es la capacidad de un instrumento PRO para detectar cambios a lo largo del tiempo en el cosntructo a medir. La respuesta está relacionada con la validez y todos los requisitos metodológicos descritos para la validez de constructo se aplican al análisis de la respuesta. La única diferencia es que la validez de constructo está relacionada con un análisis transversal (diseño clínico de corte transverso) de una puntuación única en un punto de tiempo, y la respuesta se refiere a las puntuaciones del cambio a lo largo del tiempo (diseño cohorte). En consecuencia, debe incluir información sobre hipótesis formuladas a priori del cambio en la puntuación del instrumento PRO, con la dirección y la magnitud esperada en las medias de las diferencias de la puntuación del cambio. Para escalas binarias o dicotómicas, se debe determinar la sensibilidad y especificidad (cambio versus ningún cambio). Para escalas o instrumentos PRO cuya respuesta se mide en base a una variable continua, la recomendación es calcular el tamaño del efecto (ES = media de la diferencia de puntuación del cambio dividido por la DS de la puntuación basal) o la media estandarizada de la respuesta (SRM = media de la diferencia de puntuación del cambio dividido por la DS de la puntuación del cambio). El ES se usa con frecuencia en el meta-análisis en lugar de SRM porque rara vez

se informa el denominador del SRM. Un SRM o ES grande indica una alta sensibilidad al cambio clínico; ES o SRM> 0.8 implica una mejoría clínica importante. Cuando la correlación entre las puntuaciones base y la de seguimiento tras un tiempo "t" es igual a 0.5, el ES es igual al SRM. Cuando la correlación es superior a 0,5, el SRM será mayor que el ES y cuando sea menor, el ES será mayor que el SRM (40-42). La comparación de diferentes medidas de resultado (medidas de examen clínico e instrumentos PRO) podría hacerse con el ES o SRM para evaluar cuáles son más sensibles para medir el efecto del tratamiento (41).

INTERPRETABILIDAD

La interpretabilidad no es considerada una propiedad de medición por el COSMIN, pero si una característica importante de un instrumento de medición. Se puede definir como el grado en que se puede asignar un significado clínico cualitativo a las puntuaciones cuantitativas de un instrumento o al cambio en las puntuaciones. La interpretabilidad debe ser analizada mediante el MIC o MID, la forma más común es mediante el uso de una pregunta "ancla" (mejoría vs no mejoría) y curvas de ROC para obtener el punto de corte de una puntuación de un instrumento PRO que puede ser considerada como la diferencia mínima de importancia clínica que con mayor sensibilidad y menor número de falsos positivos se ve asociada a un buen resultado o mejoría clínica. Este concepto, MIC o MID no tiene nada que ver con la MDC que está relacionada con el error de medición de in instrumento PRO (22, 23, 33).

En resumen, un instrumento PRO puede ser 100% fiable y 100% inválido. Por otro lado, un instrumento fiable y válido puede ser que no tenga respuesta para detectar cambios de importancia clínica.

REFERENCIAS

1. Kleinert HE, Kasdan ML, Romero JL. Small blood-vessel anastomosis for salvage of severely injured upper extremity. J Bone Joint Surg Am, 1963; 45: 788-96.

2. Early PF. Population studies in Dupuytren´s Contracture. J Bone Joint Surg Br, 1962; 44: 602-13.

3. Nuno-Alegre D, Azevedo L, Ferreira N, et al. Doença de Dupuytren. Revisao de 100 doentes operados. Rev Iberoamer Cir Mano, 2000; 34: 49-59.

4. Zancolli EA, Zancolli ER, Cagnone JC. Rizartrosis del pulgar. Tratamiento quirúrgico en estados iniciales y tardíos. Rev Iberoamer Cir Mano, 2000; 27: 8-18.

5. Irisarri C. Aetiology of Kienböck,s disease. J Hand Surg Br, 2004; 29: 279-85.

6. Pérez de la Fuente T, Vega García C, Alvelay Laso O, et al. Tumor de células gigantes de vainas tendinosas: una serie retrospectiva de 107 casos. Rev Iberoamer Cir Mano, 2003; 30: 29-33.

7. Rosales RS. Clinical research in hand surgery. J Hand Surg Eur , 2015 ;40:546-8.

8. Tadjerbashi K, Rosales RS, Atroshi I. Intervention randomized controlled trials involving wrist and shoulder arthroscopy: a systematic review. BMC Musculoskelet Disord. 2014, 25; 15:252.

9. Amadio PC, Higgs P, Keith M. Prospective comparative clinical trials in The Journal of Hand Surgery American, J Hand Surg Am, 1996; 21: 925-9

10. Gummeson C, Atroshi I, Ekdahl C. The quality of reporting and outcome measures in randomized clinical trials related to upper-extremity disorders. J Hand Surg Am, 2004; 29: 727-34.

11. Rosales RS, Reboso-Morales L, Martin-Hidalgo Y, et al.. Level of evidence in hand surgery. BMC Res Notes. 2012, 2; 5: 665. doi: 10.1186/1756-0500-5-665

12. World Health Organization. How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF). Exposure draft for comment. Geneva: WHO , 2013.

13. Rosales RS, Bensenny-Delagdo E, Díez de La Lastra-Bosch I. Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality of life instruments: Cross-cultural adaptation process and reliability. J Hand Surg Am, 2002; 27: 334-44.

14. Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Joint Surg Am, 1993; 75: 1585-92.

15. Atroshi I, Lyrén PE, Gummesson C. The 6-item CTS symptoms scale: a brief outcomes measure for carpal tunnel syndrome. Qual Life Res. 2009, 18:347-58.

16. Rosales RS, Atroshi I. Spanish versions of the 6-item carpal tunnel syndrome symptoms scale (CTS-6) and palmar pain scale. J Hand Surg Eur Vol. 2013 ;38:550-1.

17. Rosales RS, Martin-Hidalgo Y, Reboso-Morales L, et al. Reliability and construct validity of the Spanish version of the 6-item CTS symptoms scale for outcomes assessment in carpal tunnel syndrome. BMC Musculoskelet Disord. 2016, 3:115. doi: 10.1186/s12891-016-0963-5.

18. Beaton DE, Katz JN, Fossel AH, et al. Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. J Hand Ther, 2001; 14: 128-46.

19. Kennedy CA, Beaton DE, Solway S, et al. The DASH and Quick DASH outcome measure user´s manual. Third Edition. Toronto, Ontario: Institute for Work & Health. 2011.

20. MacDermid JC, Wessel J, Humphrey R, et al. Validity of self-report measures of pain and disability for persons who have undergone arthroplasy for osteoarthritis of the carpometacarpal joint of the hand. Osteoarthritis and Cartilage. 2007;,15 :524-530.

21. Rosales RS, García-Gutierrez R, Reboso-Morales L, et al. The Spanish version of the Patient-Rated Wrist Evaluation outcome measure: cross-cultural adaptation process, reliability, measurement error and construct validity. Health Qual Life Outcomes. 2017, 24:169. DOI: 10.1186/s12955-017-0745-2.

22. Mokkink LB, Terwee CB, Knol DL et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010 , 18:10:22. doi: 10.1186/1471-2288-10-22.

23. Terwee CB, Mokkink LB, Knol DL, et al.. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012, 21: 651-7. doi: 10.1007/s11136-011-9960-1.

24. Van Abswoude AAH, Van der Ark LA, Sijtsma K. A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. Applied Psychol Meas.2004,28:3□24.

25. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007,16 Suppl 1:5□18.

26. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951, 16: 297-334.

27. Vaz S, Falkmer T, Passmore AE, et al. The case for using the repeatability coefficient when calculating test-retest reliability. PLoS One. 2013, 8(9):e73990. doi: 10.1371/journal.pone.0073990.

28. Fleiss JL, Levin B, Paink MC. Statistical Methods for Rates and Proportions. 3rd ed. Hoboken NJ. John Wiley & Sons. 2003, 598:608.

29. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968,70:213□20.

30. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Edu Psychol Meas.1973, 33: 613-9.

31. Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1981; 1: 307-310.

32. de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006. 22:4:54.

33. Rodrigues JN, Mabvuure NT, Nikkhah D, et al. Minimal important changes and differences in elective hand surgery. J Hand Surg Eur . 2015,40:900-12.

34. Rodrigues J, Zhang W, Scammell B, et al. Validity of the Disabilities of the Arm, Shoulder and Hand patient-reported outcome measure (DASH) and the Quickdash when used in Dupuytren's disease. J Hand Surg Eur Vol. 2016 ,41:589-99.

35. Ware JE Jr, Gandek BL, Keller SD. The IQOLA Project Group. Evaluating instruments used cross-nationally: methods from the IQOLA project. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed.Philadelphia: Lippincott-Raven, 1996:337–346.

36. Rosales RS, Delgado EB, Díez de la Lastra-Bosch I. Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality-of-life instruments: cross-cultural adaptation process and reliability. J Hand Surg Am. 2002, 27:334-43.

37. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. J Clin Epidemiol. 1993, 46:1417-32.

38. Petersen MA, Groenvold M, Bjorner JB, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res. 2003,12:373□85.

39. Alexander M,. Franko OI, Makhni EC, et al. Validation of a modern activity hand survey with respect to reliability, construct and criterion validity. J Hand Surg Eur, 2008; 33: 653-60.

40. Liang MH. Evaluating Measurement Responsiveness. J Rheumtol. 1995, 22: 1191-1192.

41. Rosales RS, Diez de la Lastra I, McCabe S, Ortega Martinez JI, et al. The relative responsiveness and construct validity of the Spanish version of the DASH instrument for outcomes assessment in open carpal tunnel release. J Hand Surg Eur. 2009, 34:72-5.

42. 4 Rosales RS, Atroshi I. Basics of Statistics for Clinical Research in Hand Surgery. Rev Iberam Cir Mano. 2018, 46:141–16. DOI: 10.1055/s-0038-1675587