Grado de cumplimiento de la recomendación STARD y calidad de los estudios de precisión diagnóstica en Fisioterapia: revisión sistemática

Compliance of the STARD guideline and quality of the diagnostic validity studies in Physiotherapy: systematic review

Rismondo F[a], Ríos-Díaz J[b, c]

[a] Istituto di Medicina dello Sport di Torino-F.M.S.I. Centro di Eccellenza Federale per la Ricerca dello Sport. Torino. Italia

[b] Centro de Ciencias de la Salud San Rafael. Universidad Antonio de Nebrija. Madrid. España

[c] Fundación San Juan de Dios. Madrid. España

Correspondencia:
José Ríos-Díaz
jrios@nebrija.es

Recibido: 13 marzo 2017
Aceptado: 28 junio 2017

RESUMEN

Objetivo: evaluar la calidad de los estudios de precisión diagnóstica relacionados con el área de Fisioterapia con el listado de recomendaciones STARD y con la herramienta QUADAS-2. *Material y método:* se realizó una búsqueda en la base de datos Pubmed/Medline siguiendo las indicaciones de los autores Haynes y Leeflang, utilizando términos específicos de precisión diagnóstica (*sensitivity, specificity, diagnostic, predictive value, likelihood, ROC, Physical Therapy*) sin el uso de filtros y se mejoró su especificidad añadiendo términos propios de este tipo de estudios. Finalmente la calidad de los trabajos se evaluó con la *Quality Assessment Tool for Diagnostic Accuracy Studies* (QUADAS-2; 2010) y también con la recomendación STARD (*Standards for Reporting of Diagnostic Accuracy*). *Resultados y discusión:* se localizaron un total de 282 trabajos de los cuales se excluyeron 257. La muestra final fue de 25 artículos. La mayoría de los artículos evaluados demuestran seguir de forma incompleta las recomendaciones propuestas en QUADAS-2. En general, observando los datos obtenidos, parece que la STARD haya sido mejor aplicada que QUADAS-2. Sin embargo se evidencia la falta de una correcta aplicación metodológica en el apartado de resultados. *Conclusiones:* los ítems de la guía STARD se utilizan correctamente en la mayoría de los apartados que componen los artículos, pero se incumplen en gran medida los relacionados con la metodología y la exposición de resultados. La herramienta QUADAS-2 refleja que existe presencia de sesgo en la mayoría de los trabajos analizados.

Palabras clave: reproducibilidad de resultados, exactitud de datos, Fisioterapia, modalidades de fisioterapia, diagnóstico, QUADAS, STARD.

ABSTRACT

Objective: *to evaluate the quality of diagnostic accuracy studies related to Physiotherapy with the STARD statement and QUADAS-2 tool.* Material and method: *a research was performed with Pubmed/MEDLINE data base following guidelines by the authors Haynes and Leeflang, and using specific terms of diagnostic accuracy (sensitivity, specificity, diagnostic, predictive value, likelihood, ROC, Physical Therapy) without using filters and its specificity was improved by adding terms used in this kind of study. Finally quality of works was evaluated with the Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2; 2010) and also with the STARD initiative (Standards for Reporting of Diagnostic Accuracy)*. Results and discussion: *a total of 282 articles were located*

and 257 of them were excluded. The final sample consisted of 25 works. The majority of examined articles, erroneously follow the recommendations proposed in QUADAS-2. In general, observing obtained data, it appears that STARD have been better applied to QUADAS-2. However the lack of a proper methodological application in the results section is evidenced. Conclusions: *the items in the STARD guide are used correctly in most of the sections but not in methodology and results exposition items. The analysis with QUADAS-2 tool shows that there is a high risk of bias in most of the articles.*

Keywords: *reproducibility of results, data accuracy, physical therapy specialty, physical therapy modalities, diagnosis, QUADAS, STARD.*

INTRODUCCIÓN

En los estudios de precisión diagnóstica, la realización de un diagnóstico correcto es fundamental para una buena práctica clínica ya que pretende determinar la presencia o ausencia de una condición clínica así como la monitorización de la evolución de una enfermedad o su gravedad. Cuando se establece un diagnóstico, comienza un proceso de toma de decisiones dirigidas a establecer el pronóstico y el tratamiento de la patología[1].

Una prueba diagnóstica adecuada es aquella que proporciona resultados positivos en un paciente enfermo y negativos en un paciente sano[2], y que permite diferenciar varias condiciones clínicas que de otro modo podrían confundirse[3].

En la selección de una prueba diagnóstica pueden influir diferentes aspectos, pero son 3 los que deben tenerse en consideración: i) la «validez» en el sentido de si la prueba o el test mide aquello para lo que ha sido creado[4, 5], ii) la «reproducibilidad» entendida como la capacidad para ofrecer los mismos resultados cuando se repite su aplicación en circunstancias similares, y iii) la «seguridad» que ofrece la prueba en cuanto a tasas de acierto y de error y que vendrá determinada por la sensibilidad y especificidad[6], los valores predictivos[7], las curvas ROC[8] o, mejor aún, por los valores predictivos *(likelihood ratios*) positivo y negativo[9]. Estos aspectos que se han desarrollado históricamente bajo el paraguas de la medicina son total y necesariamente aplicables a las pruebas y test diagnósticos utilizados en Fisioterapia.

En los últimos años, se ha llamado la atención sobre las pruebas diagnósticas y su precisión en el ámbito de las ciencias de la salud[10]. Si bien se ha observado un aumento y un perfeccionamiento de las mismas, éste no ha sido acompañado por una adecuada calidad metodológica[11, 12].

En este contexto surge la recomendación *Standards for the Reporting of Diagnostic Accuracy Studies* (STARD)[13] con una estructura análoga a las ya establecidas, reconocidas por la mayoría de los comités editoriales de las revistas científicas, para el informe de ensayos clínicos (CONSORT), para estudios observacionales (STROBE) o para revisiones sistemáticas y metaanálisis (PRISMA)[14].

El listado, en su versión del 2005[11], consta de 25 ítems distribuidos en 5 secciones de las cuales las relativas a materiales y métodos y los resultados son las más importantes. Es necesario resaltar que, aunque la recomendación STARD pueda ser una buena guía para editores, revisores, autores e incluso lectores de estudios sobre precisión diagnóstica, no se trata de una escala diseñada para evaluar la calidad metodológica ni está diseñada para conocer el riesgo de sesgos en este tipo de estudios.

Para ello se publicó en 2003 una herramienta que fue concebida con este objeto, *Quality Assessment of Diagnostic Accuracy Studies* (QUADAS)[15]. Se diseñó a través de una metodología Delphi para obtener una serie de dominios encaminados a detectar la presencia del riesgo de sesgos metodológicos y la aplicabilidad de los estudios de precisión diagnóstica.

Algunos autores han explorado la implantación de la recomendación STARD en algunas revistas biomédicas españolas[12], sin embargo, hasta donde los autores han podido revisar, no existen trabajos específicos en el ámbito de la Fisioterapia.

Por tanto, los objetivos de este trabajo fueron: 1. Describir el grado de seguimiento de la recomendación STARD en los trabajos de precisión diagnóstica en el ámbito de la Fisioterapia, y 2. Evaluar la calidad metodoló-

gica, el control de los sesgos y la aplicabilidad de los trabajos de precisión diagnóstica en el ámbito de la Fisioterapia con la herramienta QUADAS-2.

MATERIAL Y MÉTODO

Diseño del estudio

Se diseñó una revisión sistemática en la que las unidades objeto de análisis fueron los trabajos sobre precisión diagnóstica en el ámbito de la Fisioterapia.

Selección de la muestra

Las búsquedas se restringieron a la base de datos Pubmed/MEDLINE con un último acceso en agosto de 2015.

La selección de los trabajos que compusieron la muestra de esta revisión se realizó según las estrategias propuestas por los autores[16–18] que diseñaron los algoritmos de las búsquedas a través de la herramienta PubMed Clinical Queries[19]. Esta estrategia proporciona una sensibilidad del 98 % y una especificidad del 74 %[20] y se completó con términos específicos para mejorar la especificidad junto a los descriptores para filtrar los trabajos del área de Fisioterapia:

#1. *sensitiv* [Title/Abstract];* ***#2.*** *sensitivity and specificity [MeSH Terms];* ***#3.*** *diagnose [Title/Abstract];* ***#4.*** *diagnosed [Title/Abstract];* ***#5.*** *diagnoses [Title/Abstract]; #6. diagnosing [Title/Abstract];* ***#7.*** *diagnosing[Title/Abstract];* ***#8.*** *diagnosis [Title/Abstract];* ***#9.*** *diagnostic [Title/Abstract];* ***#10.*** *diagnosis [MeSH:noexp];* ***#11.*** *diagnostic* [MeSH:noexp];* ***#12.*** *diagnosis,differential [MeSH:noexp];* ***#13.*** *diagnosis[Subheading:noexp];* ***#14.*** *"predictive value*"[tw];* ***#15.*** *likelihood [tw];* ***#16.*** *ROC [tw];* ***#17.*** *"ROC Curv*"[tw];* ***#18.*** *"predictive rule*"[All Fields];* ***#19.*** *Clinical Trial [ptyp];* ***#20.*** *Review[ptyp];* ***#21.*** *physiotherapy [tw];* ***#22.*** *"physical therapy"[tw];* ***#23.*** *#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 #14 OR #15 OR #16 OR #17;* ***#24.*** *#23 AND (#21 OR #22) NOT (#18 OR #19 OR #20)*

Criterios de exclusión

Tras la selección de los trabajos en la revisión manual se excluyeron todos aquellos que no se hubieran diseñado con el objetivo principal de comprobar la precisión diagnóstica de una prueba o de un test como las cartas al director, los estudios de casos o los estudios con diseño de casos y controles (que no deberían utilizarse para evaluar pruebas diagnósticas). Se excluyeron también los trabajos publicados en idiomas diferentes al español, inglés, francés, portugués o italiano.

Variables de estudio

Las variables registradas se derivan de los ítems y dimensiones de la recomendación STARD y de la escala QUADAS-2.

Para la recomendación STARD (tabla 1) se utilizó la versión publicada por Altman y Bossuyt en 2005[11]. Está compuesta por 25 ítems distribuidos que hemos considerado de forma dicotómica (Sí se cumple/No se cumple).

Para la escala QUADAS-2 (tabla 2) se adaptó la versión inglesa de la University of Bristol[15]. Está compuesta por 4 dominios: selección de pacientes, prueba diagnóstica en estudio, prueba de referencia y el flujo y secuencia temporal.

En cada dominio se desarrollan preguntas orientadas a determinar la presencia de sesgo y las dudas acerca de la aplicabilidad de la prueba. Un determinado dominio se considera como bajo riesgo de sesgo cuando todas las preguntas orientadas se contestan como bajo, en el caso de que alguna de las preguntas orientadas se puntúe como alto o dudoso se describe como riesgo de sesgo o dudas acerca de la aplicabilidad.

Todos los trabajos fueron evaluados por los dos revisores de forma independiente y cegada, y en los casos en los que hubo desacuerdo los dos revisores analizaron de forma conjunta el caso para tomar la decisión final.

Análisis de datos

Se preparó una base de datos para cada una de las herramientas y se realizó un análisis de frecuencias abso-

TABLA 1. **Listado de recomendación STARD**[11].

Sección	Ítem	Descriptor
Título/Resumen/ Palabras clave	1	Identifica el artículo como un estudio de precisión diagnóstica (epígrafe MeSH recomendado «sensibilidad y especificidad»).
Introducción	2	Especifica los objetivos del estudio, como estimación de la precisión diagnóstica o comparación de la precisión entre distintos grupos
Métodos		
Participantes	3	Población estudiada; criterios de inclusión y exclusión, contexto y centros en que se obtuvieron los datos.
	4	Proceso de selección de los participantes: ¿el proceso de selección estuvo fundamentado en los síntomas iniciales, en los resultados obtenidos en pruebas previas o en el hecho de que los participantes se hubiesen evaluado mediante la prueba índice o estándar de referencia?
	5	Selección de los participantes: ¿la población de pacientes estudiada constituyó una serie consecutiva según los criterios de selección de los puntos 3 y 4? En caso negativo, especifique el método de selección empleado.
	6	Recogida de los datos: ¿se había planificado antes (estudio prospectivo) o bien después (retrospectivo) de la realización de las pruebas índice y de referencia?
Realización de la prueba	7	Prueba estándar de referencia y su fundamento.
	8	Especificaciones de carácter técnico sobre materiales y métodos para la realización de las pruebas, incluyendo tipo, momento y referencias bibliográficas.
	9	Definición y fundamento de las unidades, valores umbral y categorías de los resultados de las pruebas índice y de referencia.
	10	Número y formación de las personas que realizan las pruebas. Coste económico.
	11	Especifique si los investigadores que interpretan todas las pruebas permanecen cegados. Describa toda información clínica relevante.
Métodos estadísticos	12	Métodos para el cálculo o la comparación de la precisión diagnóstica y para cuantificar la incertidumbre (p. ej., intervalos de confianza del 95 %).
	13	Si procede, métodos para el cálculo de la reproducibilidad y fiabilidad de la prueba.
Resultados		
Participantes	14	Fechas de realización del estudio, incluyendo las de comienzo y de finalización de la recogida de datos.
	15	Características clínicas y demográficas de los participantes en el estudio (p. ej., edad, sexo, espectro de síntomas iniciales, comorbilidad, tratamientos actuales, centros).
	16	Número de pacientes que cumplen los criterios de participación en el estudio y que se evalúan o no mediante las pruebas índice y de referencia; descripción de las razones por las que los pacientes no se evaluaron mediante alguna de estas pruebas (se recomienda el uso de un diagrama de flujo).

Resultados de la prueba	17	Intervalos de tiempo transcurridos entre la realización de las pruebas índice y estándar, así como descripción de cualquier tratamiento utilizado entre ambas.
	18	Descripción de la gravedad de la enfermedad (definir los criterios) en los pacientes que la presentan y descripción de otros diagnósticos en aquellos que no la presentan.
	19	Tabla de resultados obtenidos con la prueba índice (incluyendo los indeterminados y los no obtenidos) según los resultados de la prueba estándar. Para variables continuas, la distribución de los resultados según los valores del estándar de referencia.
	20	Todo evento adverso observado durante la realización de las pruebas.
Estimaciones	21	Estimaciones de la precisión diagnóstica y de la incertidumbre estadística (p. ej., intervalos de confianza del 95 %).
	22	Métodos para considerar los resultados extremos o no determinados y los datos ausentes.
	23	Si procede, estimación de la heterogeneidad de la precisión diagnóstica entre grupos de participantes, de investigadores o de centros.
	24	Si procede, estimación de la reproducibilidad o fiabilidad de la prueba.
Discusión	25	Discusión de la aplicabilidad clínica de los resultados del estudio.

TABLA 2. **Herramienta QUADAS-2[15].**

Descripción	**Selección de los pacientes.** Describe los métodos utilizados para seleccionar a los pacientes: pruebas previas, ámbito, uso previsto de la prueba en estudio. **Prueba diagnóstica en estudio.** Describe la prueba, cómo se realizó y su interpretación. **Prueba de referencia.** Describe la prueba de referencia, cómo se realizó y su interpretación. **Flujo y secuencia.** Describe a los pacientes que no van a recibir la prueba de estudio, la prueba de referencia o que se excluyen de la tabla 2x2: describe el intervalo y cualquier intervención entre la prueba en estudio y la de referencia.
Preguntas clave (sí/no/dudoso)	**Selección de los pacientes. 1A.1** ¿Es una muestra consecutiva o aleatoria? **1A.2** ¿Se evitó un diseño de casos y controles? **1A.3** ¿Se evitaron exclusiones inapropiadas? **Prueba diagnóstica en estudio. 1B.1** ¿Se interpretaron los resultados de la prueba sin el conocimiento de la prueba de referencia? Lo correcto es realizar primero la prueba de estudio. **1B.2** Si se usó un punto de corte (umbral), ¿se especificó previamente? **Prueba de referencia. 1C.1** ¿La prueba de referencia clasifica correctamente la enfermedad en estudio? **1C.2** ¿Los resultados de la prueba de referencia se interpretaron independientemente de la prueba de estudio? ¿Hay algún elemento de la prueba en estudio que forme parte de la prueba de referencia? **Flujo y secuencia. 1D.1** ¿Describe el intervalo de tiempo entre las dos pruebas? ¿El intervalo de tiempo es el adecuado? **1D.2** ¿Se aplicó a todos los pacientes el patrón de referencia? **1D.3** ¿Todos los pacientes recibieron la misma prueba de referencia independientemente del resultado de la prueba en estudio? **1D.4** ¿Se incluyeron todos los pacientes en el análisis?

Riesgo de sesgo (alto/bajo/ dudoso)	**Selección de los pacientes. 2A.1** ¿Hay sesgo en la selección de los pacientes? **Prueba diagnóstica en estudio. 2B.1** ¿Podría haber sesgos en la realización e interpretación de la prueba? **Prueba de referencia. 2C.1** ¿Podría haber sesgos en la realización e interpretación de la prueba? **Flujo y secuencia. 2D.1** ¿El flujo de seguimiento del paciente podría haber producido algún sesgo?
Aplicabilidad (alta/baja/dudosa)	**Selección de los pacientes. 3A.1** ¿Hay dudas de que los pacientes incluidos y su ámbito de estudio no se ajusten a la pregunta de la revisión? Es decir, que sean diferentes de la población diana. **Prueba diagnóstica en estudio. 3B.1** ¿Hay dudas de que la prueba (realización e interpretación) difieran de la pregunta de revisión? Cualquier modificación de la tecnología, interpretación o realización merma su aplicabilidad. **Prueba de referencia. 3C.1** ¿Hay dudas de que a condición de estudio (enfermedad) definida por la prueba de referencia (realización e interpretación) difiera o no se ajustara a la pregunta de revisión?

QUADAS: *Quality Assessment of Diagnostic Accuracy Studies*. Adaptada de la versión original inglesa.

lutas y relativas para cada uno de los ítems o dimensiones con los correspondientes diagramas de barras horizontales según las recomendaciones de los autores creadores de la escala. Se calcularon los valores medios (DE: desviación estándar), rango y media y cuartiles para el número de ítems cumplido en la recomendación STARD.

RESULTADOS

Identificación y proceso de selección de los estudios

Se revisaron los resúmenes y los títulos de los 282 artículos recuperados, se excluyeron 220 (78,0 %) bien porque su objetivo primario o el diseño no era el de un estudio de precisión diagnóstica, 13 por tratarse de cartas al editor (4,6 %), 7 (2,5 %) por el idioma de publicación (polaco, ruso, alemán, húngaro o danés) y 17 (6,0 %) por no estar directamente relacionados con el área de Fisioterapia (se consensuaron entre los autores). Por tanto, la muestra final quedó configurada por 25 trabajos (8,9 %) (figura 1).

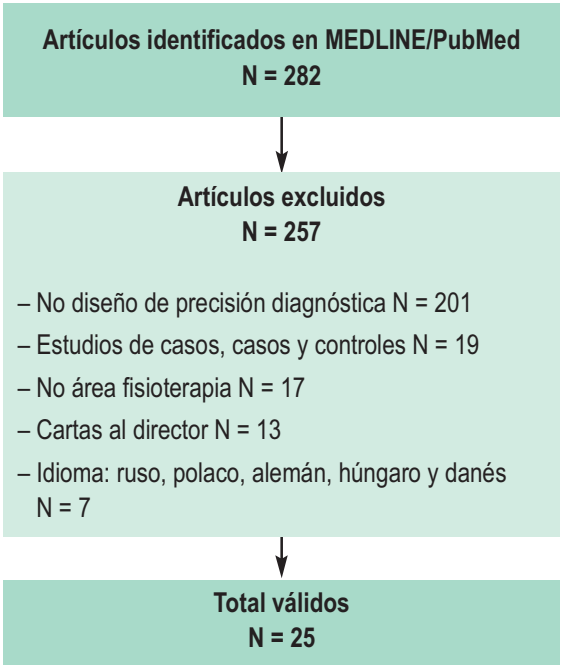


FIGURA 1. **Diagrama de flujo del proceso de búsqueda.**

Descripción de los artículos

En la tabla 3 se resumen las características de los trabajos incluidos que abarcan un período desde 1990[21, 22] hasta 2014[23, 24].

El tamaño muestral de los estudios osciló entre 20[24] y 680[25] sujetos: en 10 estudios fue superior a 100 sujetos[25–34], en 6 estudios estuvo comprendido entre 50 y 100 sujetos[21, 34–38] y en 9 trabajos fue inferior o igual a 50 sujetos[22, 24, 39–45].

TABLA 3. **Características de los trabajos incluidos en el estudio.**

Amendt[21] (1990)	**Sujetos.** 65 pacientes con escoliosis idiopática. Edad: 5 - 37 años (D.E..2,5 años). ***Index Test.*** *Scoliometer®* ***Gold Standard.*** Radiografía (ángulo de Cobb). **Resultados.** El Scoliometer® demostró una buena especificidad, sensibilidad y capacidad de predicción. Los coeficientes de fiabilidad intraevaluador e interevaluador fueron altos (r = 0,86 – 0,97). Estos datos indican una buena reproducibilidad. Sin embargo, el *Scoliometer®* por si solo, no es suficiente para determinar un diagnóstico y su gestión.
Cooperman[49] (1990)	**Sujetos.** 32 pacientes con dolor en una rodilla. Edad: 18 - 57 años; edad media: 26 años (D.E. 1,5 años). **Grupo Control.** Rodilla sana de los mismos pacientes. ***Index Test.*** Test de Lachman (para determinar una lesión del LCA). **Resultados.** Los valores Kappa intraevaluadorr fueron: 0,44 para los fisioterapeutas, 0,60 para los cirujanos ortopédicos y 0,51 para todos los examinadores. Los valores Kappa Interevaluador fueron: 0,69 para los fisioterapeutas, 0,61 para los cirujanos y 0,42 para todos los examinadores. La predicción de verdaderos positivos fue del 47 % para todos los examinadores, mientras que el valor predictivo de verdaderos negativos fue del 70 %. Los resultados indican que el Test de Lachman es más útil para predecir la ausencia de una lesión del LCA, que para la predicción de una afectación del LCA.
LaStayo[45] (1995)	**Sujetos.** 50 pacientes con dolor de muñeca. Edad: 16 - 67 años; edad media 38 años (D.E. 2,3 años). ***Index Test.*** *Scaphoid Shift Test* (SST); *Ballottement Test* (BALLOT); *Ulnomeniscotriquetral Dorsal Glide Test* (UMTDG). ***Gold Standard.*** Artroscopia. **Resultados.** Sensibilidad: SST 69 %, BALLOT 64 %, UMTDG 66 %. Especificidad: SST 66 %, BALLOT 44 %, UMTDG 64 %. Valores predictivos positivos: SST 48 %, BALLOT 24 %, UMTDG 58 %. Valores predictivos negativos: SST 78 %, BALLOT 81 %, UMTDG 69 %.
Strand[34] (1999)	**Sujetos.** 337 pacientes con dolor musculo esquelético. Edad: 21 - 64 años; edad media: 43 años. (D.E. 3,4 años). ***Index Test.*** *Sock Test, Norwegian Pain Questionnaire* (NPQ), *Disability Rating Index* (DRI). ***Gold Standard.*** Evaluaciones previas y examen físico. **Resultados.** Los resultados del *Sock Test* estaban relacionados con los informes de la limitación de la actividad diaria. El aumento de la edad y del índice de masa corporal incrementó la probabilidad de resultados que evidencian una limitación en las actividades. Los valores pretest fueron predictivos de dificultades a la hora de vestirse después de 1 año.
Fritz[35] (2000)	**Sujetos.** 69 pacientes con dolor lumbar. Edad media: 37,3 años (D.E. 1,2 años). ***Index Test.*** *Nonorganic symptom, Questionnaires,* examen físico. **Resultados.** Tras 4 semanas de tratamiento, 47 pacientes volvieron al trabajo normalmente, mientras los restantes 22 pacientes volvieron al trabajo con restricciones. Se calcularon la sensibilidad, especificidad y *likelihood ratio* para los valores obtenidos en los signos, síntomas e índice (signos y síntomas). La *likelihood ratio* negativa para los signos fue 0,75; para los síntomas fue 0,62 y para el índice fue 0,59.

Holtby[56] (2004)	**Sujetos.** 50 pacientes con afectación del manguito de los rotadores y sometidos a cirugía de hombro. Edad: 24 - 79 años; edad media: 50 años (D.E. 2,8 años). ***Index Test.*** *Supraspinatus Test.* ***Gold Standard.*** Artoscopia, cirugía. **Resultados.** Se evaluó la sensibilidad, la especificidad y la *likelihood ratio* de 3 grupos de pacientes en función de la gravedad de la lesión: 1. Tendinitis o rotura parcial del tendón del supraespinoso; 2. Rotura completa del tendón del supraespinoso; 3. Rotura completa del supraespinoso con afectación de otros tendones del manguito de los rotadores (infraespinoso, redondo menor y subescapular). La sensibilidad de la prueba del supraespinoso fue, en los tres grupos, del 62, 41 y 88 % respectivamente. Los valores de especificidad fueron del 54, 70 y 70 % respectivamente. Los valores *likelihood ratio* en los casos negativos variaron de 0,17 a 0,84, en los casos positivos variaron de 1,35 a 2,93 dependiendo de la presencia de dolor o debilidad.
Kim[57] (2004)	**Sujetos.** 48 pacientes (54 hombros) con inestabilidad postero inferior del hombro, sin dolor. Edad: 19 - 31 años; edad media: 24 años (D.E. 0,9 años). **Grupo Control.** 33 pacientes (35 hombros) con inestabilidad postero inferior del hombro, con dolor. Edad: 19 - 29 años; edad media: 25 años. ***Index Test.*** *Shoulder Jerk Test,* escala EVA (para el dolor y la función), *Rowe Score, UCLA scale, American Shoulder and Elbow Surgeons* (ASES). ***Gold Standard.*** Resonancia magnética. **Resultados.** En el grupo sin dolor, 50 hombros (93 %) respondieron al programa de rehabilitación después de una media de 4 meses. Cuatro hombros (7 %) resultaron insensibles a la rehabilitación. El grupo control tuvo una mayor tasa de fracaso con el tratamiento no quirúrgico ($P < 0,001$): 5 hombros (16 %) tuvieron éxito con la rehabilitación, mientras que los otros 30 hombros (84 %) falló. Todos los 34 hombros que no respondían a la rehabilitación tenían un grado variable de lesión en el labrum posteroinferior. El *Shoulder Jerk Test* es un sello distintivo para predecir el pronóstico del tratamiento no quirúrgico para la inestabilidad posteroinferior.
Shaw[33] (2004)	**Sujetos.** 105 pacientes con posible disfagia. Edad: 17 - 96 años; edad media: 71 años (D.E. 3,7 años). ***Index Test.*** Auscultación bronquial (BA). ***Gold Standard.*** Videofluoroscopia (VF). **Resultados.** La comparación se hizo entre los resultados de la VF equipo y los resultados de la BA en la clasificación de los pacientes como «aspiración» o en «riesgo de aspiración». Un alto grado de consenso se encontró en el caso de «riesgo de aspiración» (sensibilidad 87 %), aunque la especificidad fue baja (37 %). BA fue altamente específico (88 %) cuando se confirma la ausencia de aspiración, pero la sensibilidad a la presencia de aspiración era 45 %.
Laslett[32] (2005)	**Sujetos.** 107 pacientes con dolor lumbar. Edad media: 42,9 años (D.E. 1,2 años). ***Index Test.*** Examen físico con método McKenzie, *Roland–Morris Disability Questionnaire*, escala EVA, *Modified Somatic Perception Questionnaire* (MSPQ), *Distress Risk Assessment Method* (DRAM). ***Gold Standard.*** Tomografía computarizada (TC). **Resultados.** En relación con una tomografía positiva, las técnicas de evaluación del método McKenzie tienen una especificidad del 89 %, y entre los pacientes sin dolor intenso es del 100 %. Sin embargo, en presencia de discapacidad grave, la especificidad se reduce al 80 %. El estudio sugiere que en los pacientes con un leve dolor lumbar, se puede retrasar la TC, si está disponible un programa de tratamiento McKenzie, ya que además existe un buen pronóstico con tratamiento conservador.

Laslett[31] (2006)	**Sujetos.** 120 pacientes con dolor lumbar. Edad media: 43 años (D.E. 1,7 años). ***Index Test.*** Examen físico y tratamiento con el método McKenzie, *Roland–Morris Disability Questionnaire*, escala EVA, *Zung Depression Index, Modified Somatic Perception Questionnaire* (MSPQ), *Distress Risk Assessment Method* (DRAM). ***Gold Standard.*** Tomografía computarizada (TC). **Resultados.** El artículo ha sido redactado de forma incomprensible.
Feick[36] (2008)	**Sujetos.** 87 pacientes con sospecha de hidrocefalia normotensiva. Grupo Respondedor: Edad media 72,3 años (D.E. 9,0 años). Grupo no Respondedor: Edad media 72,0 años (D.E. 13,3 años). **Grupo Control.** 33 pacientes que no respondían al tratamiento de drenaje del líquido cefalorraquídeo. ***Index Test.*** *Functional Independence Measure* (FIM), *Timed Up & Go Test* (TUG), *Tinetti Assessment Tool of Gait and Balance* (Tinetti), *9-Hole Peg Test* (Peg Test), Cognitive Assessment of Minnesota (CAM), *Mini Mental State* Exam (MMSE). **Resultados.** Todas las variables, medidas tras el drenaje del líquido cefalorraquídeo en los pacientes que respondían positivamente a este tratamiento, mostraron una notable mejoría (52 % *vs* 11%). Las evaluaciones llevadas a cabos en el ámbito de la terapia ocupacional y de la fisioterapia evidenciaron una buena sensibilidad al cambio y un interesante valor predictivo.
Frost[30] (2008)	**Sujetos.** 201 pacientes con dolor de espalda. Edad media: 42,5 años (D.E. 1,4 años). ***Index Test.*** *Patient Specific Activity Questionnaire* (PSAQ), *Oswestry Disability Index versión 2.1* (ODI), *Roland and Morris Disability Questionnaire* (RMDQ), *Global Transition Rating Scale* (utilizada solamente para dividir en 3 grupos los pacientes: grupo en el que hubo una mejoría del dolor, grupo en el que el dolor permaneció igual y grupo en el que hubo un empeoramiento). **Resultados.** Todos los instrumentos fueron capaces de detectar mejoras en el dolor de espalda. En el grupo de la mejoría los resultados del cuestionario PSAQ fueron 1,08 – 1,31 y del ODI fueron de -0,88 a -1,00; sin embargo para el RMDQ fueron limitados (de -0,70 a -0,74). En el grupo que refería un empeoramiento del dolor, los valores fueron: ODI (0,61-1,16), RMDQ (0,69 a 1,25), PSAQ (de -0,16 a -0,26).
Woodley[42] (2008)	**Sujetos.** 40 pacientes con dolor de cadera. Edad media: 54,4 años (D.E. 2,2 años). ***Index Test.*** Examen clínico. ***Gold Standard.*** Resonancia magnética. **Resultados.** Patología del tendón del glúteo medio, bursitis, osteoartritis y atrofia del glúteo menor fueron nombrados en todos los informes relativos a la resonancia magnética. Algunas anormalidades eran identificadas también en las caderas asintomáticas, como por ejemplo la bursitis. Sobre el diagnóstico, no hubo mucho acuerdo entre los radiólogos y tampoco con los fisioterapeutas.
Raney[37] (2009)	**Sujetos.** 68 pacientes con dolor cervical. Edad media: 47,8 años (D.E. 2,7 años). **Index Test.** Tracción cervical, ejercicio de fortalecimiento. **Resultados.** Treinta pacientes (44 %) fueron clasificados como haber logrado con éxito el tratamiento, mientras los 38 restantes no obtuvieron un resultado exitoso. En el primer grupo los valores obtenidos para el dolor con la escala NPRS fueron 2,2, 95 % CI = 1,2 – 3,2, mientras en la discapacidad fueron 12,5, 95 % CI = 6,2 – 18,7.
Iqbal[58] (2010)	**Sujetos.** 124 casos de pacientes con lesión del rodete glenoideo. Edad: 20 - 54 años; edad media: 37 años (D.E. 1,7 años). ***Index Test.*** Artrografía. ***Gold Standard.*** Artroscopia. **Resultados.** Sólo 51 pacientes fueron sometidos a artroscopia. En la artrografía hubo 4 falsos positivos (15,4 %) y un falso negativo (3,84 %). La sensibilidad fue del 95,6 % (22/23), la especificidad del 85,7 % (24/28), el valor predictivo positivo del 84,6 % (22/26) y el valor predictivo negativo fue del 96 % (24/25). La artrografía es una técnica útil para el diagnóstico y la planificación preoperatoria cuando hay la sospecha de una lesión del rodete glenoideo. También puede evitar artroscopias innecesarias.

Farrell[40] (2011)	**Sujetos.** 34 pacientes con demencia. Edad media: 76,6 años (D.E. 3,1 años). ***Index Test.*** *Physical performance test* (PPT) (incorpora múltiples dominios de evaluación incluyendo las AVD, motricidad gruesa, control motor, equilibrio y marcha). **Resultados.** Se estudiaron tres variables (caídas anteriores a la medición, puntuación del PPT y edad), sólo una caída en los 6 meses previos a la medición fue un predictor de una caída en los 4 meses de seguimiento (p = 0,044). Las probabilidades de caída aumentaron de casi 5 veces (95 % CI = 1,04 – 21,8). La sensibilidad fue del 58%, la especificidad fue del 77 %. La *likelihood ratio* en los casos negativos fue de 0,58 y en los casos positivos la proporción fue de 2,52.
Kott[29] (2011)	**Sujetos.** 440 pacientes. Edad: 3 - 18 años. ***Index Test.*** *Standardized Walking Obstacle Course* (SWOC). **Resultados.** La edad, el peso, y la discapacidad fueron los predictores más importantes (p < 0,05). El aumento de la edad y el peso predijeron un tiempo más corto y menor cantidad de pasos durante la realización del test. Por otro lado la discapacidad predijo un tiempo más largo y un aumento de los pasos.
Strand[59] (2011)	**Sujetos.** 98 pacientes con dolor de espalda. Edad: 18 - 60 años; Edad media: 37,4 años (D.E. 1,1 años). ***Index Test.*** Tests de las funciones corporales: *Biering-Sørensen test, Spondylometry DeBrunner kyphometer* (para la medición de la movilidad espinal sagital), *Global Physiotherapy Examination, Lift Test, Lateral flexion test*, distancia dedo-suelo. Tests de las actividades: *Progressive Isoinertial Lifting Evaluation*, 15-m (50-ft) *Walk Test, Back Performance Scale. Hannover Functional Ability Questionnaire, Roland-Morris Disability Questionnaire* (RMDQ). **Resultados.** Cinco pruebas físicas demostraron una buena capacidad de respuesta (sensibilidad al cambio): *Spondylometry, Lateral flexion test*, distancia dedo-suelo, *Lift Test, Back Performance Scale.*
Bohman[25] (2012)	**Sujetos.** 680 pacientes con latigazo cervical y sintomatología añadida (WAD). Edad media: 39 años (D.E. 1,0 años). ***Index Test.*** Formulario rellenado por el paciente (está compuesto por tres dominios de interés: salud general y estado psicológico, socio-demográfico y comorbilidad). **Resultados.** La capacidad predictiva (c-index) aumentó en los tres dominios y alcanzó un nivel aceptable de 0,68 (IC del 95 %: 0,65, 0,71). La validez interna (c-index: 0,67 (95 % IC: 0,63, 0,70)) evidenció una capacidad aceptable del test para predecir la recuperación del paciente.
Brach[27] (2012)	**Sujetos.** 552 pacientes. Edad media: 79,4 años (D.E. 1,5 años). ***Index Test.*** *Stance Time Variability Test (STV), Mobility Disability* (test de auto evaluación). **Resultados.** Un año después de la primera medición, 59 pacientes indicaron dificultad para caminar. En la primera medición, los resultados del test STV fueron: sensibilidad 65 %, especificidad 65 % y área de la curva ROC 0,70. Un año más tarde los resultados del test STV fueron: sensibilidad 61 %, especificidad 60 % y área de la curva ROC 0,65.
Garvey[26] (2012)	**Sujetos.** 158 pacientes con dolor inguinal crónico. Edad: 18 - 88 años; edad media: 43 años (D.E. 3,4 años). ***Index Test.*** Tomografia computarizada (TC). ***Gold Standard.*** Examen físico. **Resultados.** Los pacientes con diagnóstico de hernia fueron: 45 verdaderos positivos y 4 falsos positivos. Los pacientes sin diagnóstico de hernia fueron: 94 verdaderos negativos y 5 falsos negativos. En el preoperatorio, el diagnóstico hecho con la TC, tuvo un valor predictivo positivo (VPP) del 92 % y un valor predictivo negativo (VPN) del 96 % (precisión global 94 %). La TC puede ser un complemento útil a la evaluación de los pacientes que presentan un dolor inguinal sin diagnosticar.

Cunningham[41] (2013)	**Sujetos.** 29 pacientes con dolor lumbar. Grupo derivado al médico: edad mediana 66,5 años (RI: 10 años). Grupo no derivado al médico: edad mediana 62,0 años (RI: 23 años). ***Index Test.*** *Screen Assist Lumbar Questionnaire* (SALQ) (para recopilar información subjetiva relevante sobre el paciente y determinar los casos que requieren la derivación a otro especialista). ***Gold Standard.*** Diagnóstico hecho por el médico de atención primaria. **Resultados.** El SALQ demostró una sensibilidad del 100 % (95 % CI = 0,44 – 1,0) y especificidad del 92 % (95 % CI = 0,81 – 0,92). La *likelihood ratio* en los casos negativos fue de 0,11 (95 % CI = 0,01 – 1,54) y en los casos positivos la proporción fue de 9,36 (95 % CI = 2,78 - 32). Si el SALQ fue positivo, la probabilidad postest fue 0,60. Si el SALQ fue negativo, la probabilidad postest fue 0,017. Los resultados de este estudio sugieren que el SALQ se puede utilizar como complemento en el reconocimiento de los trastornos musculoesqueléticos emergentes que requieren la derivación a otro profesional.
Karel[38] (2013)	**Sujetos.** 60 pacientes con dolor de hombro. Edad: ≥ 18 años. **Grupo Control.** 60 pacientes con dolor de hombro. ***Index Test.*** Examen físico, tratamiento de fisioterapia, *Global Perceived Effect Test* (GPE), *Shoulder Disability Questionnaire* (SDQ-NL), *Shoulder Pain Disability Index* (SPADI), *Shoulder Pain Score* (SPS), *Euroquol* 5 ítems (EQ-5D), *Working Alliance Inventory* (WAV-12). **Gold Standard.** Ecografía músculo esquelética, resonancia magnética, rayos-X. **Resultados.** Los fisioterapeutas cada vez más utilizan la ecografía músculo esquelética en la práctica diaria y la fiabilidad entre las diferentes profesiones aún no ha sido evaluado. Se espera que este estudio ayude a mejorar la gestión actual y el pronóstico de los pacientes con dolor en el hombro.
Christy[24] (2014)	**Sujetos.** 20 pacientes hipoacusia neurosensorial bilateral severa o profunda con una edad media 8,9 años (D.E. 1,8 años). **Grupo Control.** 23 niños con desarrollo normal con edad media de 9,5 años (D.E. 2,9 años). ***Index Test.*** *Head Thrust Test, Emory Clinical Vestibular Chair Test, Bucket Test, Dynamic Visual Acuity, Modified Clinical Test of Sensory Interaction on Balance, Sensory Organization Test.* ***Gold Standard.*** *Vestibular Evoked Myogenic Potential Test y Rotary Chair Test.* **Resultados.** La fiabilidad osciló entre un coeficiente de correlación intraclase de 0,73-0,95. Los marcadores de sensibilidad, especificidad y valores predictivos, oscilaron entre 63 y 100 %. Las pruebas clínicas de este estudio podrán ser utilizadas con precisión para identificar a los niños con hipofunción vestibular.
Vance[39] (2015)	**Sujetos.** 36 pacientes con Parkinson. Edad media: 71,4 años (D.E.1,6 años). Grupo 1: 19 pacientes sin caídas previas. **Grupo Control.** Grupo 2: pacientes con caídas en los anteriores 6 meses. ***Index Test.*** *Timed Up & Go Test* (TUG). **Resultados.** Se clasificaron tres tipos de test TUG: estándar, cognitivo y manual. El TUG cognitivo mostró un óptimo rendimiento discriminativo (área bajo la curva = 0,82; 95 % IC = 0,64 – 0,92). El TUG cognitivo era más probable que clasificara correctamente a los participantes con un bajo riesgo de caída (*likelihood ratio* positiva = 2,9) y obtuvo los valores más altos de sensibilidad (0,76; 95 % IC = 0,52 – 0,90) que los de especificidad (0,73; 95 % IC = 0,51 – 0,88) en este umbral (*likelihood ratio* negativa = 0,32).

Grado de cumplimiento del listado STARD

El cumplimiento medio fue de 14,4 ítems (D.E: 3,74) con un rango entre 3[30] y 20 ítems[24, 44]. La mediana fue de 15 ítems, con un 25 % de los trabajos con menos de 12 ítems y un 25 % con más de 17.

Los ítems correspondientes a «Título/resumen, Introducción y Discusión» tienen un alto porcentaje de cumplimiento, entre el 72 y el 80 % (figura 2).

Para la sección de «métodos» (figura 2), que comprende 11 ítems (desde el 3 al 13) se cumplen por la mayoría de los trabajos excepto el ítem 9 (no lo cumple el 32 %), el ítem 10 (no lo cumple el 100 %), el ítem 11 (no lo cumple el 52 %) y el ítem 13 (no lo cumple el 72 %).

Finalmente, en el área relativa a los «resultados» (11 ítems desde el 14 al 24), se encontró que los ítems 14, 17, 20, 23 y 24 mostraron un elevado grado de incumplimiento (superior al 80 %) tal como se refleja en la figura 2.

Calidad de los estudios según la escala QUADAS-2

En la tabla 4 se muestran las codificaciones y las puntuaciones obtenidas para cada trabajo en cada una de las preguntas dirigidas y los cuatro dominios de la escala QUADAS-2 (figura 3).

Para el primer dominio «Método de selección» 5 trabajos (20 %) presentaban un riesgo de sesgo, 7 (28 %) dudoso y solamente 5 (20 %) una aplicabilidad clara.

En el segundo dominio, «Prueba de estudio» (*Index test)*, se encontró que en 2 (8 %) de los trabajos existía un riesgo de sesgo claro y en 11 (44 %) un sesgo probable. Solamente en 3 (12 %) la aplicabilidad era clara.

En el tercer dominio, relacionado con la «Prueba de referencia», solamente 7 trabajos (28 %) no mostraban riesgo de sesgo y los restantes 18 (72 %) o lo presentaban claramente o de forma dudosa. Once (44 %) de los trabajos no mostraron una aplicabilidad para este dominio.

Finalmente, en el dominio «Flujo y secuencia» se encontró que el riesgo de sesgo aparecía en 9 (36 %) de los estudios y de forma probable en 12 (48 %).

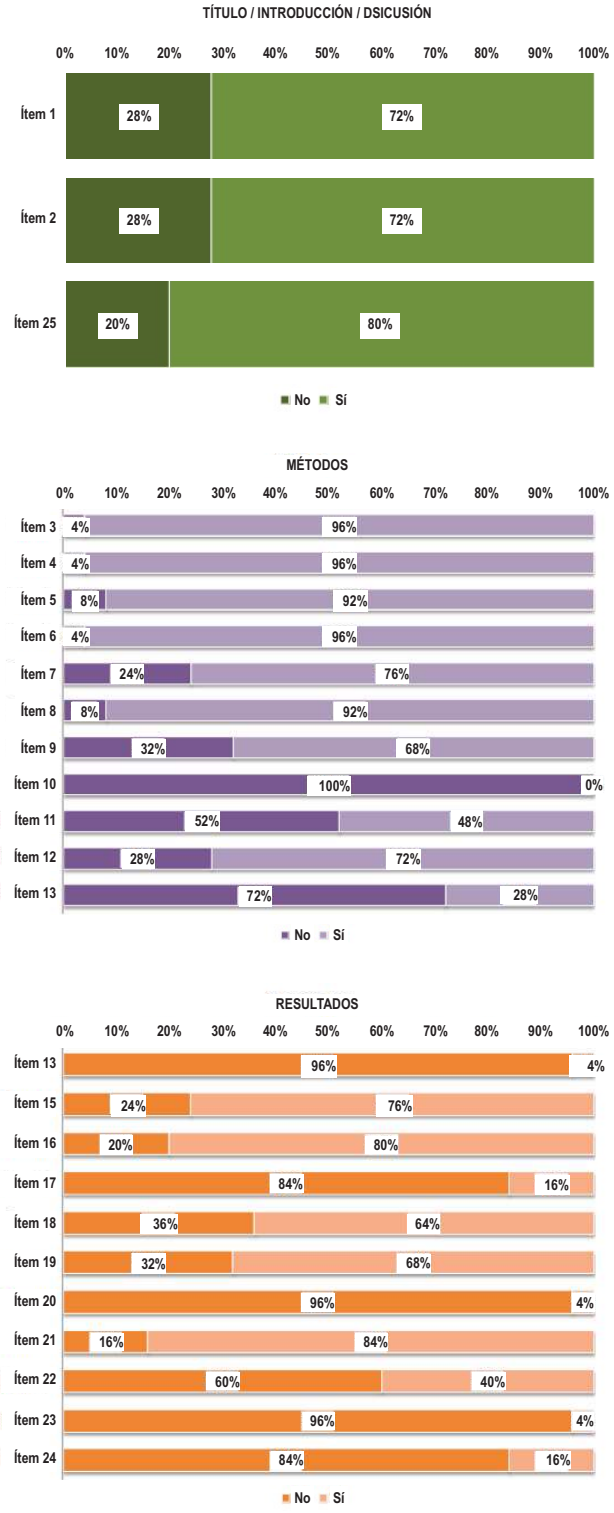


FIGURA 2. **Diagrama de barras apiladas para el grado de cumplimiento de la lista STARD.**

TABLA 4. **Matriz de datos para la escala QUADAS-2.**

Artículo	Selección de los pacientes					Prueba de diagnóstico en estudio				Prueba de referencia				Flujo y cronograma				
	Preguntas clave			Riesgo de Sesgo	Aplica-bilidad	Preguntas clave		Riesgo de Sesgo	Aplica-bilidad	Preguntas clave		Riesgo de Sesgo	Aplica-bilidad	Preguntas clave				Riesgo de Sesgo
	1A.1	1A.2	1A.3	2A.1	3A.1	1B.1	1B.2	2B.1	3B.1	1C.1	1C.2	2C.1	3C.1	1D.1	1D.2	1D.3	1D.4	2D.1
Amendt[21] (1990)	2	1	0	1	1	2	1	2	0	2	2	2	0	0	1	1	1	1
Brach[27] (2012)	1	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
Bohman[25] (2012)	1	1	1	0	0	1	1	0	0	2	2	2	2	2	2	2	2	2
Christy[24] (2014)	2	2	0	1	1	1	1	0	0	1	2	2	2	2	0	1	0	1
Cooperman[49] (1990)	1	1	1	0	0	1	2	2	2	1	1	0	0	1	1	1	1	0
Cunningham[41] (2013)	1	1	1	0	0	1	1	0	0	1	1	0	0	2	1	1	0	1
Farrell[40] (2011)	2	2	1	2	2	2	1	2	2	1	2	2	0	2	1	1	1	2
Feick[36] (2008)	1	1	0	1	1	1	1	0	0	2	2	2	2	2	1	1	1	2
Fritz[35] (2000)	1	1	1	0	0	2	1	2	2	1	2	2	2	0	1	1	1	1
Frost[30] (2008)	1	1	0	1	1	2	0	1	1	0	2	1	1	2	0	0	1	1
Garvey[26] (2012)	1	1	1	0	0	1	1	0	0	2	2	2	2	2	2	2	1	2
Holtby[56] (2004)	1	1	1	0	0	1	1	0	0	1	1	0	0	2	0	0	1	1
Iqbal[58] (2010)	2	1	1	0	0	1	1	0	2	1	2	1	0	0	0	0	1	1

Artículo	Selección de los pacientes					Prueba de diagnóstico en estudio				Prueba de referencia				Flujo y cronograma				
	Preguntas clave			Riesgo de Sesgo	Aplicabilidad	Preguntas clave		Riesgo de Sesgo	Aplicabilidad	Preguntas clave		Riesgo de Sesgo	Aplicabilidad	Preguntas clave				Riesgo de Sesgo
	1A.1	1A.2	1A.3	2A.1	3A.1	1B.1	1B.2	2B.1	3B.1	1C.1	1C.2	2C.1	3C.1	1D.1	1D.2	1D.3	1D.4	2D.1
Karel[38] (2013)	1	1	1	0	0	1	1	0	2	1	1	0	0	2	2	2	1	2
Kim[57] (2004)	1	1	1	0	2	2	1	2	2	1	2	2	0	2	1	1	1	2
Kott[29] (2011)	0	2	1	1	1	2	1	2	2	2	2	2	2	2	1	1	1	2
Laslett[32] (2005)	1	1	2	2	2	1	1	0	0	1	1	0	0	1	1	1	1	0
Laslett[31] (2006)	2	2	1	2	2	1	1	0	0	1	2	2	2	2	1	1	1	1
Raney[37] (2009)	2	2	1	2	2	0	1	1	1	0	0	1	1	0	0	0	0	1
Shaw[33] (2004)	1	1	1	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0
Strand[60] (1999)	1	1	1	0	0	2	2	2	2	1	2	2	2	1	1	1	1	0
Strand[59] (2011)	1	1	1	0	0	2	2	2	1	2	2	2	1	2	2	2	2	2
Vance[23] (2015)	1	1	1	0	0	2	1	2	2	2	2	2	2	1	1	1	2	2
Woodley[42] (2008)	2	2	1	2	2	1	1	0	0	1	1	0	0	2	1	2	1	2
LaStayo[45] (1995)	2	2	1	2	2	2	1	2	0	1	2	2	2	2	1	1	1	2

El número 1 seguido de la letra corresponde con las preguntas guiadas. El numero 2 seguido de la letra con el riesgo de sesgo y el número 3 seguido de la letra con la aplicabilidad. Las letras desde la A hasta la D se refieren a cada una de los cuatro dominios. La codificación se estableció para las preguntas guiadas como 0: no cumple criterio, 1=si cumple criterio, 2=dudoso. Para los sesgos 0=no riesgo de sesgo, 1=riesgo de sesgo, 2= dudoso. Para la aplicabilidad 0=no aplicable, 1=aplicable, 2=dudoso.

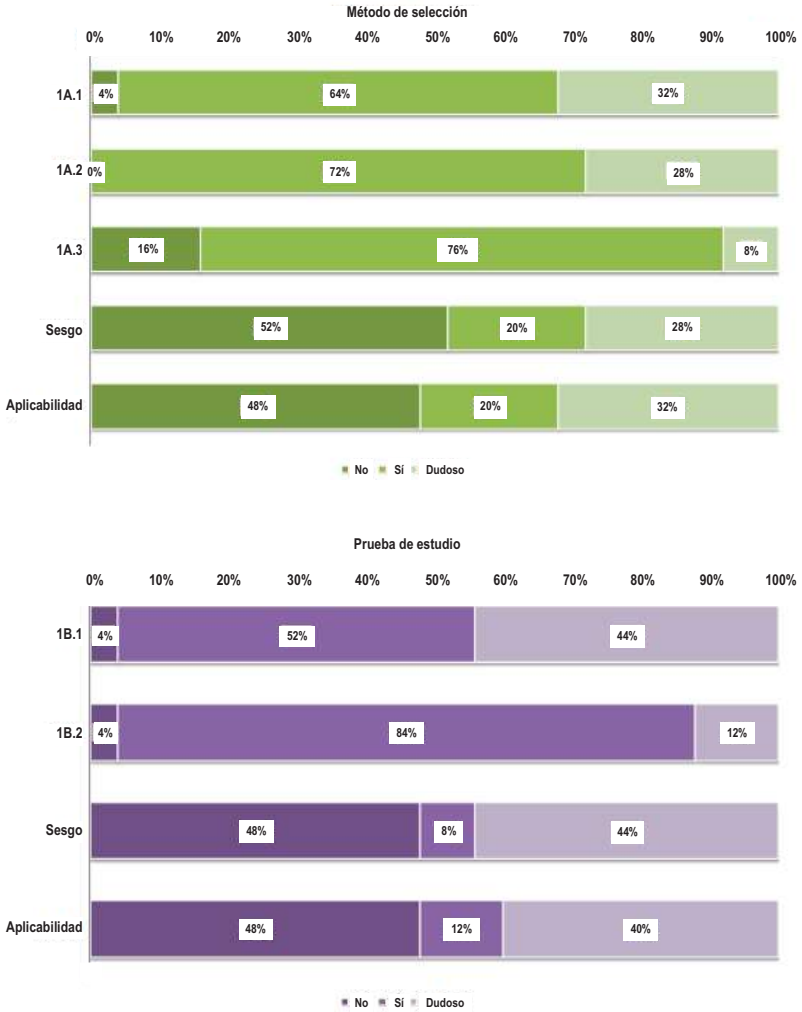
DISCUSIÓN

Uno de los objetivos de este trabajo fue evaluar el grado de seguimiento del listado STARD de los estudios de precisión diagnóstica en el área de la Fisioterapia. Se ha podido comprobar que aunque determinados ítems se identificaban en la mayoría de los trabajos, los porcentajes de cumplimiento eran deficientes en algunos de los aspectos más críticos, tanto del método como de la exposición de los resultados.

Por ejemplo, el ítem 11 que hace referencia a si se describen métodos de cegado de los investigadores, es incumplido por la mitad de los trabajos analizados y el ítem 12 que refiere a los métodos estadísticos es incumplido por una tercera parte. El ítem 13 que se refiere a la descripción de la fiabilidad de la prueba, solamente lo cumple una tercera parte de los trabajos estudiados.

Con la publicación de la declaración STARD[46] en 2003, se pretendía mejorar la comunicación de los trabajos con diseño de precisión diagnóstica; sin embargo, diferentes estudios han mostrado que el éxito ha sido relativamente bajo. En el área de Fisioterapia, esta revisión es la primera aproximación al grado de cumplimiento de la recomendación STARD.

Smidt y cols.[47] revisaron los trabajos sobre precisión diagnóstica publicados en 12 revistas biomédicas, ninguna de ellas del área de fisioterapia, en el año 2000 (pre-STARD) y en 2004 (post-STARD) y encontraron un ligero aumento en el valor medio de ítems cumplidos, en especial los relativos al cálculo de la reproducibilidad de



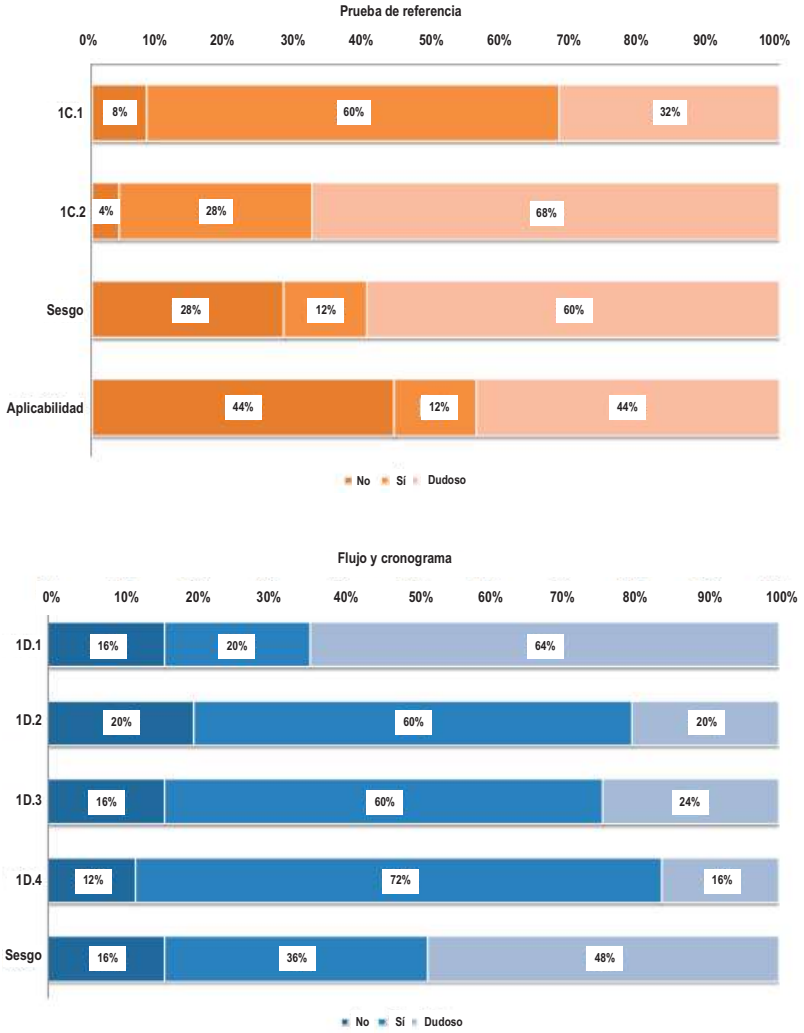


FIGURA 3. **Diagrama de barras para los dominios de QUADAS-2.**

la prueba índice, la distribución de la gravedad de la enfermedad y otros diagnósticos, la estimación de la variabilidad de la precisión diagnóstica entre los subgrupos y la presencia de un diagrama de flujo.

Sin embargo, Wilczynsk y cols.[48] comprobaron el grado de mejoría de la calidad de las publicaciones sobre precisión diagnóstica después de la publicación de la declaración STARD y si existían diferencias entre las revistas que la habían incorporado a sus recomendaciones editoriales y las que no. Los resultados mostraron el cumplimiento de los criterios se mantuvieron constantes antes y después de la publicación de STARD y que existía una leve diferencia no significativa en la calidad de los informes entre las revistas que recomendaban su uso y las que no. Tampoco encontraron una mejoría significativa en el grado de cumplimiento a lo largo de los 4 años analizados (2001-2005). Aunque no era el objetivo de nuestro estudio, cuando se observan las puntuaciones finales en relación al año de publicación de los trabajos, tampoco se observa una clara asociación.

Coppus y cols.[49] también evaluaron el impacto de la declaración STARD en el ámbito de la medicina reproductiva entre los años 1999 y 2004 y concluyeron la inadecuada descripción del diseño, de la metodología y del

análisis estadístico a pesar de la publicación de la declaración.

En el ámbito nacional, sólo se ha encontrado el trabajo de Gómez y cols. (2009)[12] en el que comprobaron el grado de cumplimiento en los trabajos publicados en cuatro de las revistas nacionales de mayor impacto (Medicina Clínica (Barc), Enfermedades Infecciosas y Microbiología Clínica, Revista Española de Cardiología y Archivos de Bronconeumología) en el período 2004-2007. El grado de cumplimiento fue de 11,9 ítems (D.E. 3,7 ítems), inferior a los 14,4 ítems (D.E. 3,7 ítems) de nuestro estudio. No obstante, el porcentaje de cumplimiento en los diferentes ítems fue similar al encontrado en nuestro estudio.

Tan sólo se ha encontrado un trabajo que evalúe la adherencia a la declaración STARD en un área relacionada con la Fisioterapia. En 2006, Rama y cols.[50] revisaron la calidad de los informes de los estudios de precisión diagnóstica en las tres mayores revistas de ortopedia (*Clinical Orthopaedics and Related Research, Journal of Bone and Joint Surgery British Volume, and Journal of Bone and Joint Surgery American Volume*) en el período 2002-2004. Las puntuaciones oscilaron entre 6,6 y 21,4 con una media de 15 (D.E. 3,3) ítems cumplidos, puntuaciones similares a los obtenidos en la presente revisión (media 14,4; D.E. 3,7). Solamente un 38 % de los trabajos revisados cumplían más de dos tercios de los ítems y la mayoría fallaban en el cumplimiento de los nuevos ítems específicos con lo que concluían que la adhesión a los criterios STARD estaba por debajo de lo deseable. En este sentido, los resultados de nuestro estudio son ligeramente mejores puesto que el 50 % de los trabajos cumplían al menos 15 de los 25 ítems. Sin embargo, en la adhesión a criterios clave para la correcta interpretación, como se ha anotado más arriba, las tasas de cumplimiento fueron bajas.

Finalmente, Korevaar y cols.[51] publicaron un reciente metaanálisis sobre los estudios que analizan la adherencia a la declaración STARD de los estudios de precisión diagnóstica. Detectaron un modesto y significativo incremento medio de 1,4 ítems IC 95 %: 0,65 a 2,18, en la adherencia tras la introducción de la declaración STARD.

Debe tenerse en cuenta que el incumplimiento de los criterios del listado no necesariamente significa que los errores o sesgos estén presentes en los estudios. Es posible que los investigadores hayan diseñado correctamente el estudio pero no lo reflejen adecuadamente en el artículo; esto supone que el lector no tenga acceso a aspectos metodológicos relevantes para la interpretación y la lectura crítica. Además, la replicación de los estudios se ve afectada, cuando no imposibilitada, por la falta de información.

Aunque en el presente trabajo se ha utilizado la declaración STARD de 2003 por estar ampliamente extendida y para facilitar la comparabilidad con otros estudios, en 2015 se publicó una actualización con 30 ítems esenciales (frente a los 25 de la primera versión) y que deberían ser incluidos en cualquier informe de estudios de precisión diagnóstica[52]. La actualización incorpora evidencia actualizada sobre las fuentes de sesgo y variabilidad en la precisión diagnóstica e intenta facilitar aún más el uso de la declaración STARD.

En relación al segundo objetivo, el análisis de los estudios seleccionados con la herramienta QUADAS-2, que permite detectar el riesgo de sesgo y la aplicabilidad de los estudios de precisión diagnóstica seleccionados, muestra también y en general unos resultados pobres y la presencia de riesgo de sesgo alto en los cuatro dominios.

A pesar de los esfuerzos dirigidos por distintos grupos de trabajo para alcanzar una mayor calidad metodológica en el campo del diagnóstico clínico, los resultados encontrados confirman el no cumplimento de los ítems aconsejados en la mayoría de los artículos localizados, con bajas calidades metodológicas numerosas limitaciones a la hora de interpretarlos[53].

Si los diferentes estudios que han puesto de manifiesto la escasa mejoría en el desarrollo de aspectos metodológicos claves tras la implantación de la norma STARD, ampliamente difundida y divulgada desde 2003, en el caso de la herramienta QUADAS-2, la situación es aún peor como demuestra la escasez de publicaciones al respecto. Es posible que la escasa utilización de esta escala pueda deberse a la forma en que está construida y a que requiere una cierta práctica en su interpretación.

Hemos identificado 2 posibles limitaciones en la presente revisión. La primera podría ser el método de selección de los trabajos puesto que únicamente se ha utilizado la base de datos Pubmed/Medline. La locali-

zación de estudios diagnósticos presenta dificultades añadidas porque la terminología es muy variada y no siempre se utiliza con el mismo significado. A pesar de que el tesauro de MEDLINE contiene un descriptor denominado «*Sensibility and Specificity*» y dos filtros en «tipo de estudio» que aparentemente podrían resultar de utilidad, *Evaluation Studies[ptyp] OR Validation Studies[ptyp]*[54], se ha puesto de manifiesto la poca eficacia y efectividad de las búsquedas sobre estudios diagnósticos al emplear estas estrategias[16, 53]. Los términos *positive predictive value* incluyen las palabras *sensitivity, specificity, accuracy* y *precision*[16]. Por este motivo, una búsqueda con solamente los vocablos «especificidad» y «sensibilidad» no sería completa y asimismo en la mayoría de las bases de datos se utilizan de forma incongruente lo que complica la localización de trabajos. Además, la información sobre la exactitud de la prueba de diagnóstico puede estar oculta en los estudios que no tienen esta estimación como objetivo principal[55].

Sin embargo, los artículos seleccionados mediante las estrategias de búsqueda específicas y diseñadas para localizar trabajos sobre precisión diagnóstica tienen una gran sensibilidad[16–18, 53, 55]. Además, el hecho de haber escogido un descriptor general «fisioterapia» nos ha permitido acceder a una mayor heterogeneidad de trabajos por lo que pensamos que la muestra de artículos seleccionados puede considerarse una buena representación.

La segunda limitación está relacionada con la versión seleccionada de la declaración STARD (2003). Son dos los motivos: la mayoría de los trabajos publicados hasta la fecha que revisan la adherencia lo hacen sobre este listado y por tanto son más fáciles las comparaciones con otros estudios; y la actualización de 2015 no estaba completamente difundida cuando se publicaron los trabajos incluidos en la revisión.

Aunque la comunidad científica parece ser consciente de la necesidad de seguir mejorando la transparencia de los informes que se publican, se debe hacer un esfuerzo aún mayor en el uso de las guías de recomendación por parte de los investigadores, editores y revisores.

Podría ser eficaz que los autores de los estudios de precisión diagnóstica en el ámbito de la Fisioterapia tomaran la declaración STARD e incluso la herramienta QUADAS-2 como referencias desde el mismo momento en que se empiece a diseñar el estudio y no sólo en el momento de elaborar el informe escrito. De esta forma se pondría la atención en aquellos aspectos que pudieran ser fuentes de sesgo. La mejoría de la calidad posibilitará un adecuado uso de las diferentes pruebas diagnósticas en el ámbito clínico y evitaría la introducción de pruebas diagnósticas no suficientemente evaluadas o que podrían dar lugar a decisiones erróneas.

CONCLUSIONES

Los ítems de la guía STARD se utilizan correctamente en la mayoría de los apartados que componen los artículos, aunque hay un alto grado de incumplimiento en ítems relacionados con la metodología y la exposición de resultados.

El análisis de los estudios seleccionados con la herramienta QUADAS-2 muestra unos resultados pobres y la presencia de riesgo de sesgo alto en la descripción del método de selección de los sujetos, la descripción de la prueba de estudio y de la prueba de referencia, así como en la secuencia y flujo del diseño del estudio.

La utilidad de la declaración STARD e incluso a la herramienta QUADAS-2 podría ser mayor si los investigadores que realizan estudios de precisión diagnóstica en el ámbito de la Fisioterapia, las adoptasen como referencia desde la fase del diseño del estudio y no sólo en el momento de elaborar el informe escrito; de esta forma se pondría más atención en aquellos aspectos que pudieran ser fuentes de sesgo.

La mejoría de la calidad de este tipo de estudios posibilitaría un adecuado uso de las diferentes pruebas diagnósticas en el ámbito clínico y evitaría la introducción de pruebas diagnósticas no suficientemente evaluadas o que podrían dar lugar a decisiones erróneas.

RESPONSABILIDADES ÉTICAS

Protección de personas y animales. Para esta investigación no se han realizado experimentos en seres humanos.

Confidencialidad y consentimiento informado. Para esta investigación no se han realizado intervenciones en seres humanos.

Privacidad. En este artículo no aparecen datos de pacientes.

Financiación. El estudio no ha recibido ninguna clase de financiación, ni pública ni privada, para la realización del mismo.

Conflicto de interés. Los autores declaran no tener conflicto de intereses.

Autoría. Todos los autores declaran cumplir los criterios de autoría según las siguientes contribuciones: Federica Rismondo ha participado en el diseño, revisión y redacción final del manuscrito, José Ríos-Díaz coordinó, dirigió, tuteló y finalmente revisó y colaboró en la redacción del manuscrito final. Todos los autores han revisado críticamente el artículo hasta la aprobación de la versión final para su publicación.

BIBLIOGRAFÍA

1. Abraira V, Zamora J. Criterios de calidad de los estudios sobre pruebas diagnósticas. FMC Aten Primaria. 2008; 15(7): 460–1.
2. Ortín E, Sánchez JA, Menárguez JF, Hidalgo IM. Lectura crítica de un artículo sobre diagnóstico. En: Sánchez JA, coord. Atención sanitaria basada en la evidencia: Su aplicación a la práctica clínica. Murcia: Consejería de Sanidad de la Región de Murcia; 2007. p. 233–73.
3. Salech F, Mery V, Larrondo F, Rada G. Estudios que evalúan un test diagnóstico: interpretando sus resultados. Rev Médica Chile. 2008; 136(9): 1203–8.
4. Casas J, Repullo JR, Donado J. La encuesta como técnica de investigación. Elaboración de cuestionarios y tratamiento estadístico de los datos (II). Aten Primaria. 2003; 31(9): 592–600.
5. Pita S, Pértega S. Pruebas diagnósticas: Sensibilidad y especificidad. Cad Aten Primaria. 2003; 10(2): 120–4.
6. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994; 308(6943): 1552.
7. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994; 309(6947): 102.
8. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994; 309(6948): 188.
9. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004; 329(7458): 168–9.
10. Fritz JM, Wainner RS. Examining Diagnostic Tests: An Evidence-Based Perspective. Phys Ther. 2001; 81(9): 1546–64.
11. Altman DG, Bossuyt PM, STARD group, REMARK group. Diagnostic (STARD) and prognostic (REMARK) studies. Med Clínica. 2005; 125(Suppl 1): 49–55.
12. Gómez N, Hernández-Aguado I, Lumbreras B. Estudio observacional: evaluación de la calidad metodológica de la investigación diagnóstica en España tras la publicación de la guía STARD. Med Clínica. 2009; 133(8): 302–10.
13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Ann Intern Med. 2003; 138(1): W1–12
14. The EQUATOR Network. Enhancing the QUAlity and Transparency Of Health Research [Internet]. [citado 3 enero 2017]. Disponible en:http://www.equator-network.org/
15. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011; 155(8): 529–36.
16. Haynes RB. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004; 328(7447): 1040.
17. Kastner M, Wilczynski NL, McKibbon AK, Garg AX, Haynes RB. Diagnostic test systematic reviews: bibliographic search filters («Clinical Queries») for diagnostic accuracy studies perform well. J Clin Epidemiol. 2009; 62(9): 974–81.
18. Lokker C, Haynes RB, Wilczynski NL, McKibbon KA, Walter SD. Retrieval of diagnostic and treatment studies for clinical use through PubMed and PubMed's Clinical Queries filters. J Am Med Inform Assoc. 2011; 18(5): 652–9.
19. PubMed Clinical Queries [Internet]. [citado 2 de enero de 2017]. Disponible en: https://www.ncbi.nlm.nih.gov/pubmed/clinical
20. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. PubMed Help [Internet]. National Center for Biotechnology Information (US); 2016 [citado 2 de enero de

2017]. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK3827/
21. Amendt LE, Ause-Ellias KL, Eybers JL, Wadsworth CT, Nielsen DH, Weinstein SL. Validity and reliability testing of the Scoliometer. Phys Ther. 1990; 70(2): 108–17.
22. Cooperman JM, Riddle DL, Rothstein JM. Reliability and validity of judgments of the integrity of the anterior cruciate ligament of the knee using the Lachman's test. Phys Ther. 1990; 70(4): 225–33.
23. Vance RC, Healy DG, Galvin R, French HP. Dual tasking with the timed «up & go» test improves detection of risk of falls in people with Parkinson disease. Phys Ther. 2015; 95(1): 95–102.
24. Christy JB, Payne J, Azuero A, Formby C. Reliability and diagnostic accuracy of clinical tests of vestibular function for children. Pediatr Phys Ther Off Publ Sect Pediatr Am Phys Ther Assoc. 2014; 26(2): 180–9.
25. Bohman T, Côté P, Boyle E, Cassidy JD, Carroll LJ, Skillgate E. Prognosis of patients with whiplash-associated disorders consulting physiotherapy: development of a predictive model for recovery. BMC Musculoskelet Disord. 2012; 13: 264.
26. Garvey JFW. Computed tomography scan diagnosis of occult groin hernia. Hernia J Hernias Abdom Wall Surg. 2012; 16(3): 307–14.
27. Brach JS, Wert D, VanSwearingen JM, Newman AB, Studenski SA. Use of stance time variability for predicting mobility disability in community-dwelling older persons: a prospective study. J Geriatr Phys Ther 2001. 2012; 35(3): 112–7.
28. Iqbal HJ, Rani S, Mahmood A, Brownson P, Aniq H. Diagnostic value of MR arthrogram in SLAP lesions of the shoulder. Surg J R Coll Surg Edinb Irel. 2010; 8(6): 303–9.
29. Kott KM, Held SL, Giles EF, Franjoine MR. Predictors of Standardized Walking Obstacle Course outcome measures in children with and without developmental disabilities. Pediatr Phys Ther Off Publ Sect Pediatr Am Phys Ther Assoc. 2011; 23(4): 365–73.
30. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine. 2008; 33(22): 2450–7.
31. Laslett M, McDonald B, Aprill CN, Tropp H, Oberg B. Clinical predictors of screening lumbar zygapophyseal joint blocks: development of clinical prediction rules. Spine J Off J North Am Spine Soc. 2006; 6(4): 370–9.
32. Laslett M, Oberg B, Aprill CN, McDonald B. Centralization as a predictor of provocation discography results in chronic low back pain, and the influence of disability and distress on diagnostic power. Spine J Off J North Am Spine Soc. 2005; 5(4): 370–80.
33. Shaw JL, Sharpe S, Dyson SE, Pownall S, Walters S, Saul C, et al. Bronchial auscultation: an effective adjunct to speech and language therapy bedside assessment when detecting dysphagia and aspiration? Dysphagia. 2004; 19(4): 211–8.
34. Strand LI, Wie SL. The Sock Test for evaluating activity limitation in patients with musculoskeletal pain. Phys Ther. 1999; 79(2): 136–45.
35. Fritz JM, Wainner RS, Hicks GE. The use of nonorganic signs and symptoms as a screening tool for return-to-work in patients with acute low back pain. Spine. 2000; 25(15): 1925–31.
36. Feick D, Sickmond J, Liu L, Metellus P, Williams M, Rigamonti D, et al. Sensitivity and predictive value of occupational and physical therapy assessments in the functional evaluation of patients with suspected normal pressure hydrocephalus. J Rehabil Med. 2008; 40(9): 715–20.
37. Raney NH, Petersen EJ, Smith TA, Cowan JE, Rendeiro DG, Deyle GD, et al. Development of a clinical prediction rule to identify patients with neck pain likely to benefit from cervical traction and exercise. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc. 2009; 18(3): 382–91.
38. Karel YHJM, Scholten-Peeters WGM, Thoomes-de Graaf M, Duijn E, Ottenheijm RPG, van den Borne MPJ, et al. Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study. BMC Musculoskelet Disord. 2013; 14: 62.
39. Vance RC, Healy DG, Galvin R, French HP. Dual tasking with the timed «up & go» test improves detection of risk of falls in people with Parkinson disease. Phys Ther. 2015; 95(1): 95–102.
40. Farrell MK, Rutt RA, Lusardi MM, Williams AK. Are scores on the physical performance test useful in determination of risk of future falls in individuals with dementia? J Geriatr Phys Ther 2001. 2011; 34(2): 57–63.
41. Cunningham S. Diagnostic accuracy: sensitivity and specificity of the ScreenAssist Lumbar Questionnaire in com-

parison with primary care provider tests and measures of low back pain: a pilot study. J Man Manip Ther. 2013; 21(1): 48–59.

42. Woodley SJ, Nicholson HD, Livingstone V, Doyle TC, Meikle GR, Macintosh JE, et al. Lateral hip pain: findings from magnetic resonance imaging and clinical examination. J Orthop Sports Phys Ther. 2008; 38(6): 313–28.
43. Kim S-H, Park J-C, Park J-S, Oh I. Painful jerk test: a predictor of success in nonoperative treatment of posteroinferior instability of the shoulder. Am J Sports Med. 2004; 32(8): 1849–55.
44. Holtby R, Razmjou H. Validity of the supraspinatus test as a single clinical test in diagnosing patients with rotator cuff pathology. J Orthop Sports Phys Ther. 2004; 34(4): 194–200.
45. LaStayo P, Howell J. Clinical provocative tests used in evaluating wrist pain: a descriptive study. J Hand Ther Off J Am Soc Hand Ther. 1995; 8(1): 10–7.
24. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Clin Chem. 2003; 49(1): 7–18.
47. Smidt N, Rutjes AWS, Van der Windt D, Ostelo R, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement Has it improved? Neurology. 2006; 67(5): 792–7.
48. Wilczynski NL. Quality of Reporting of Diagnostic Accuracy Studies: No Change Since STARD Statement Publication—Before-and-after Study. Radiology. 2008; 248(3): 817–23.
49. Coppus SFPJ, van der Veen F, Bossuyt PMM, Mol BWJ. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. Fertil Steril. 2006; 86(5): 1321–9.
50. Rama KRBS, Poovali S, Apsingi S. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. Clin Orthop. 2006; 447: 237–46.
51. Korevaar DA, van Enst WA, Spijker R, Bossuyt PMM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med. 2014; 19(2): 47–54.
52. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015; 351: h5527.
53. Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev. 2013; 2: 82.
54. Sensitivity and Specificity - MeSH - NCBI [Internet]. [citado 2 de enero de 2017]. Disponible en: https://www.ncbi.nlm.nih.gov/mesh/68012680
55. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008; 149(12): 889–97.
56. Holtby R, Razmjou H. Validity of the supraspinatus test as a single clinical test in diagnosing patients with rotator cuff pathology. J Orthop Sports Phys Ther. 2004; 34(4): 194–200.
57. Kim S-H, Park J-C, Park J-S, Oh I. Painful jerk test: a predictor of success in nonoperative treatment of posteroinferior instability of the shoulder. Am J Sports Med. 2004; 32(8): 1849–55.
58. Iqbal HJ, Rani S, Mahmood A, Brownson P, Aniq H. Diagnostic value of MR arthrogram in SLAP lesions of the shoulder. Surg J R Coll Surg Edinb Irel. 2010; 8(6): 303–9.
59. Strand LI, Anderson B, Lygren H, Skouen JS, Ostelo R, Magnussen LH. Responsiveness to change of 10 physical tests used for patients with back pain. Phys Ther. 2011; 91(3): 404–15.
60. Strand LI, Wie SL. The Sock Test for evaluating activity limitation in patients with musculoskeletal pain. Phys Ther. 1999; 79(2): 136–45.