ARTÍCULO DE REVISIÓN

Escalas y listas de evaluación de la calidad de estudios científicos

Evaluation lists and scales for the quality of scientific studies

MSc. Franciele Cascaes da Silva, Téc. Beatriz Angélica Valdivia Arancibia, MSc. Rodrigo da Rosa Iop, Dr. Paulo Jose Barbosa Gutierres Filho, Dr. Rudney da Silva

Centro de Ciencias de la Salud y el Deporte de la Universidad del estado de Santa Catarina, Brasil.

RESUMEN

El objetivo de este estudio fue identificar las escalas de evaluación de la calidad metodológica de artículos científicos y de las listas de verificación de la calidad de información en investigaciones en el área de la salud. Fueron desarrollados dos procedimientos básicos: a) revisión sistemática de la literatura científica en las bases de datos Web of Science, Journals@ovid, Science Direct, Scopus, SportDiscus, Mary Ann Liebert y Oxford Journals Online, con selección de artículos publicados en los últimos cinco años e indexados en la lengua inglesa; b) revisión bibliométrica en las referencias de los artículos seleccionados en la revisión sistemática, sin definición del tiempo ou de la lengua. Fueron seleccionados diferentes estudios que representaron 14 escalas y sus modificaciones, y también 11 listas utilizadas. Se puede concluir que las escalas y las listas difieren entre sí en relación con el número de *items*, validad, fiabilidad y márgenes de puntuación. La mayoría de las escalas e *items* presentan propiedades psicométricas de validez y fiabilidad, y son aplicables a los estudios revisionales, principalmente metanaliticos, tanto en la búsqueda de la calidad metodológica como en la calidad de información.

Palabras clave: escalas, evaluación de la investigación en salud, metodología, información.

ABSTRACT

The objective of this study was to identify the scales of the methodological quality assessment of scientific articles and checklists of information quality in the area of

health. It was performed two basic procedures: a) systematic review of the scientific literature available in the databases Web of Science, Journals@ovid, Science Direct, Scopus, SportDiscus, Mary Ann Liebert Journals Online and Oxford, with selection of the articles published in the last five years and indexed in English; b) bibliometric review in the articles previously selected in the systematic review, without setting time or Language. We selected studies that accounted for 14 scales and their modifications, plus 11 checklists. Therefore, one can conclude that the scales and lists differ in the number of items, validity, reliability and score range and most have valid and reliable psychometric properties. It was also found that these scales and checklists are applicable to empirical studies, especially randomized controlled trials, and revisional studies, principally meta-analysis, both in pursuit of methodological quality as the quality of information.

Key words: scales, evaluation in health research, methodology, information

INTRODUCCIÓN

La evaluación de la calidad de los estudios científicos puede ser considerada esencial en el proceso de producción y selección de la literatura cientifica en la salud. Según *Verhagen*,[1] la evaluación de la calidad metodológica considera la validez interna, que hace referencia al análisis de la capacidad de mensurar adecuadamente lo que fue propuesto, y la validez externa, que se refiere al análisis de las hipótesis estadísticas y a la generalización de los resultados para la población de interés; además, permite analizar la transparencia en la descripción de los objetivos, la importancia del tamaño de la muestra para detectar el efecto clínico investigado y la presentación de los resultados.[2,3] Esta evaluación puede ser realizada por las listas de verificación y por las escalas de evaluación. Las listas de verificación son útiles cuando proporcionan orientaciones e informaciones que deben ser incluidas en los relatos de los ensayos clínicos aleatorizados (ECAs). Las escalas de evaluación proveen un índice cuantitativo de la calidad metodológica de los ECAs[4] y tienen la ventaja de que pueden ser fácilmente replicadas e incorporadas formalmente en la revisión, por medio de pares y en comentarios sistemáticos; pero también tiene desventajas como la falta de pruebas que deciden la inclusión y la exclusión de los *ítems* y sus puntajes numéricos unidos a cada uno de los elementos evaluados.

Las escalas y las listas de verificación incluyen *items* que miden la calidad de los estudios. Las escalas proporcionan respuestas para los *items* individuales que son procesados y pueden ofrecer puntajes globales que proveen puntos que clasifican la calidad metodológica, como por ejemplo, la escala de Jadad, donde esta es la más utilizada en los ECAs, que proporciona un puntaje que clasifica el estudio como débil (0 puntos) a bueno (5 puntos); o la escala PEDro, que es aplicada en estudios experimentales y que puntúa conforme a la presencia de indicadores de la calidad de la evidencia presentada (1 punto) o la ausencia de esos indicadores (0 puntos), hasta un puntaje total de 10 puntos. Se debe destacar que un único puntaje de calidad sugiere facilitación en la interpretación, en tanto algunas directrices deben ser seguidas para la evaluación del desempeño adecuado en las pruebas psicométricas de fiabilidad, validez de contenido, de constructo y validez concurrente, entre otros, que son consideradas esenciales en el proceso de calificación de la literatura científica actual.[2,3]

En un estudio realizado en la década de 1990, fueron identificadas 25 escalas de evaluación de la calidad de los estudios primarios, pero apenas solo una escala había desarrollado siguiendo los procedimientos metodológicos consolidados.[4] Actualmente, diversas escalas y listas han sido producidas con la finalidad de aumentar la calidad metodológica y de información de diferentes tipos de investigación en el área de la salud, tales como estudios clínicos no aleatorios, de observación y de revisión sistemática. Ese aumento ha proporcionado importantes herramientas para investigadores, editores científicos y lectores de diferentes áreas científicas, ya que la identificación de las escalas válidas y confiables sobre un tópico especifico puede minimizar las chances de errores en la determinación de la calidad de la literatura científica,[5] en la ejecución de un estudio y en la verificación de la aplicación de los resultados.[3,6]

Sin embargo, muchas dificultades propias de los problemas en las investigaciones en el área de la salud aún son reveladas en el proceso de la evaluación de la calidad de los estudios, como por ejemplo, las dificultades que cotidianamente los parceristas de periódicos científicos se encuentran con las incongruencias entre los objetivos, procedimientos y resultados, que en parte tienen un origen en bases teóricas poco consistentes o inadecuadamente seleccionadas; o las dificultades encontradas por analistas éticos que, a pesar del aparente cuidado de la dignidad humana propuesta por los investigadores, terminan confrontándose con preocupaciones económicas de los financiadores del estudio, como por ejemplo, la definición de la muestra, el uso indiscriminado del placebo o del *wash-out,* las debilidades en la bases teóricas de los racionales, y también las fallas empíricas de las fases preclínicas que antecedieron el estudio.

Considerando todas estas dificultades, diversas acciones internacionales han investido fuertemente en los procesos de calificación de la literatura científica en salud y en las áreas afines, tales como las *Minimum Information about a Microarray Experiment* (MIAME) y el *Minimum Information for Biological and Biomedical Investigation*, (MIBBI)[7,8]. En las investigaciones en el área de la salud, los requisitos de uniformidad para manuscritos enviados a revistas biomédicas pueden ser considerados una de las principales acciones en la búsqueda de la calificación de las publicaciones para los autores, editores, analistas y publicadores científicos[9], así como, el *Enhancing the Quality and Transparency of Health Research* (EQUATOR)[7], que reúne investigadores, editores, especialistas en metodología de la investigación y otros interesados en mejorar la calidad y la transparencia de las publicaciones por medio de directrices que ayudan a mejorar los aspectos experimentales y los resultados.

Considerando que la evaluación de la calidad metodológica y de información permite el análisis de la ejecución y de la aplicación de una investigación[3,6] que puede calificar la produción cientifica, en este artículo se buscó identificar las escalas de evaluación de la calidad metodológica de los artículos científicos y de las listas de verificación de la calidad de información en investigaciones en el área de la salud, a partir de sus características básicas referentes al número de *items*, las propiedades psicométricas de validez y fiabilidad, aplicaciones y limitaciones.

MÉTODOS

Esta revisión sistemática fue realizada de acuerdo con las recomendaciones de la Colaboración Cochrane.[10,11] Los estudios fueron buscados en las bases de datos *Web of Science*, *Journals@ovid, Science Direct, Scopus, SportDiscus, Mary Ann*

Liebert y *Oxford Journals Online* según los descriptores de los *Medical Subject Headings* (MeSH) y Descriptores de la Salud (DECs): a) listas (*checklist*), b) escala (*scales*); c) evaluación crítica de la metodología (*critical appraisal of methodology),* d) evaluación de la calidad (*quality assessment*), disponibles en las palabras clave de los artículos. La identificación, manipulación y control de las referencias bibliográficas y de los archivos fueron realizados con el *EndNote* (versión 3.5).

IDENTIFICACIÓN DE LOS ESTUDIOS

Fueron realizados dos procedimientos para identificar los estudios: a) revisión sistemática; b) revisión bibliométrica. En la revisión sistemática se identificaron los estudios a partir de criterios de inclusión que visaban la obtención de artículos completos, con titulo indexado en la lengua inglesa, producidos entre los años 2007 y 2012, provenientes del área de la salud. Los criterios de exclusión fueron utilizados para descartar los artículos que no presentaban informaciones suficientes, principalmente cuanto a sus propiedades psicométricas. La búsqueda fue estandarizada y realizada por dos revisores independientes (FCS y RS) que procedieron inicialmente a la lectura de los títulos, después de los resúmenes, y finalmente del artículo integral. Los artículos reconocidos después de la lectura integral fueron recuperados y en caso de divergencias en la obtención de los estudios, los procedimientos fueron inversamente repetidos por ambos revisores hasta que fueran corregidas las discrepancias.

SELECCIÓN DE LOS ESTUDIOS

Después de la identificación y selección primaria de ocho artículos, fue realizada una revisión bibliométrica, buscando identificar autores/obras de referencia sobre las temáticas en cuestión, y que no tuvieran establecido el tiempo o la lengua de publicación. Con la búsqueda
manual de las obras disponibles en las referencias de los artículos seleccionados se identificó la existencia de 51 estudios sobre la utilización, desarrollo y propiedades psicométricas de escalas y listas de evaluación de la calidad metodológica y de información. De estos 59 estudios, 25 artículos fueron seleccionados por tratarse del desarrollo de escalas y listas de evaluación de la calidad metodológica y de información (Fig. 1). Las principales razones para la exclusión en la revisión sistematica y en la revisión bibliométrica fueron: a) estudios muy específicos que no podrian ser aplicados en las áreas de la salud; b) estudios de garantía de la calidad; c) estudios duplicados y d) estudios de aplicación, pero no de desarrollo, de escalas y listas de evaluación de la calidad metodológica y de información.

EXTRACCIÓN Y ANÁLISIS DE LOS DATOS

La extracción de la información de los artículos seleccionados fue orientada para la obtención de las identificaciones de las escalas, de los autores, del año de publicación del artículo y consecuentemente de la escala, del número de *items* de la escala, de los índices y de las implicaciones de la validez de la escala, y de los índices e implicaciones de la fiabilidad de las escalas, así como sus aplicaciones y limitaciones.

El análisis de la información fue realizado de manera descriptiva cualitativa en relación con sus implicancias y sus límites, y cuantitativamente en relación con los índices investigados. El proceso de selección de los artículos se muestra en la figura 1. Los resultados sobre la identificación de las escalas y de las listas, de la autoría, del año de publicación, del número de *items*, de la validez y la fiabilidad de las escalas son presentados sinópticamente por medio de los cuadros 1 y 2

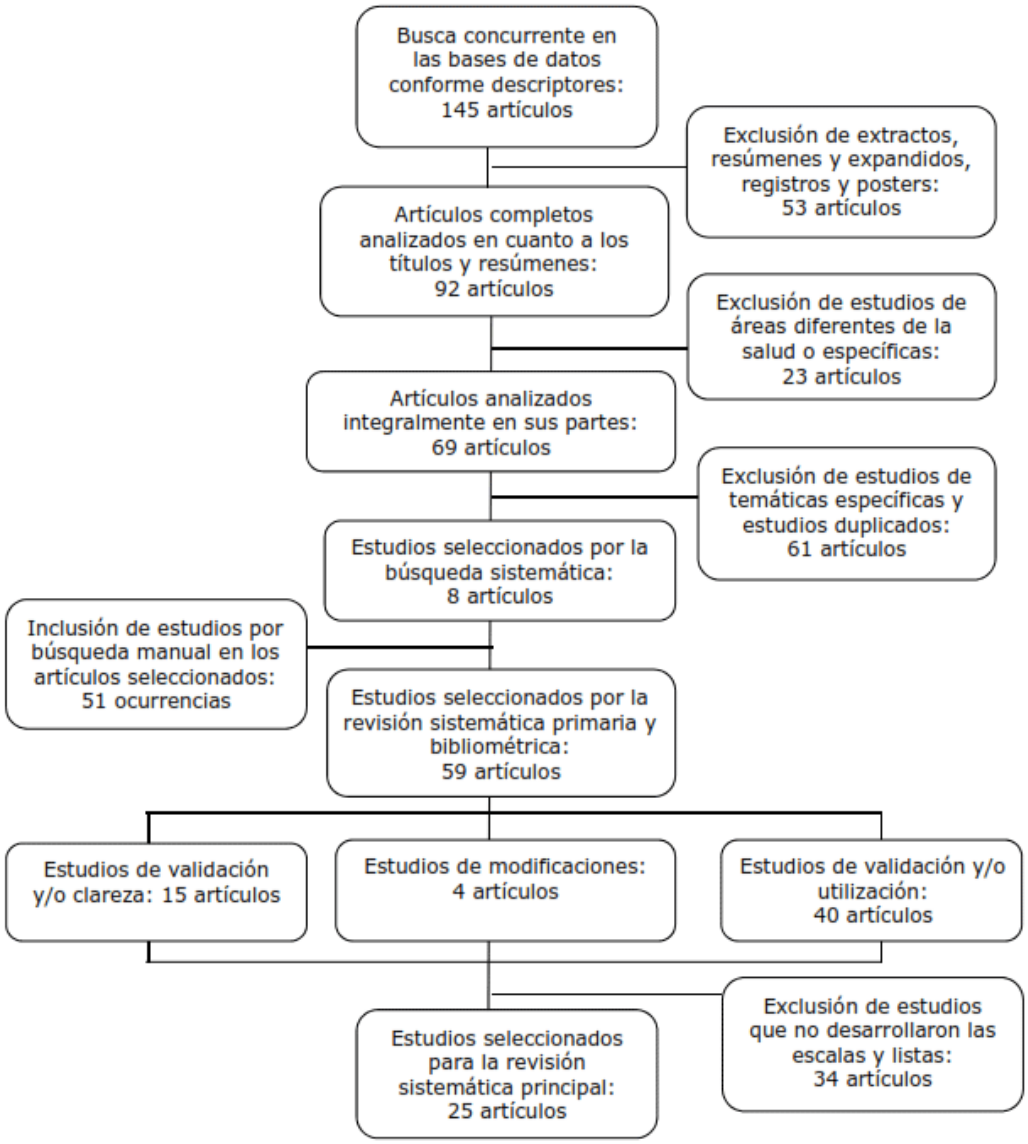

Fig. Diagrama de flujo del proceso de selección de los artículos.

RESULTADOS

Fueron identificadas un total de 14 escalas (cuadro 1) y 11 listas (cuadro 2), que ofrecen diferentes focos de evaluación y que demuestran los avances en la utilización de herramientas de evaluación de la calidad metodológica y de

información, así como también pueden ofrecer beneficios para la ciencia actual, en especial para el área de la salud.

Cuadro 1. Sinopsis de las características de las escalas de evaluación de la calidad metodológica y de la información.

Escala	Autores/año	No. de *ítems*	Validez	Confianza
Jedad	*Jedad* y otros, 1996	3 *items* seleccionados directamente relacionados con el control de errores	Validez aparente, de contenido y de constructo. Validez concurrente con la lista Delphi (Spearman r=0,63). Validez concurrente con Detsky, Imperiale, Reisch y Van Tulder fue 0,78, 0,61, 0,64 y 0,67 respectivamente	CCI varió de 0,48-1,00. A concordancia Kappa varió de 0,37 a 0,89
Maastricht	*Vet* y otros, 1997	16 *items* con base en 3 aspectos de estudio: (1) validez interna, (2) de precisión del estudio y (3) parámetros de efectos de intervención	Validez aparente. Validez concurrente con la lista Delphi y escala de Jadad fue Spearman 0,87 y 0,78 respectivamente	CCI=0,85
Single-Case Experimental Design Scale	*Tate* y otros, 2008	11 *items*	Validez de contenido	Puntuación total (CCI=0,83 y CCI=0,88). Para todos los *items* K=0,48-1,00
Van Tulder	*Van Tulder* y otros, 2003	11 *items*	Validez aparente y de contenido. Validez concurrente con las escalas Detsky, Imperiale, Jadad y Reisch fue de 0,89, 0,75, 0,67 y 0,77 respectivamente	k=0,66, k=0,29, k= 0,42, k=0,67, k=0,65, k=0,74
PEDro	*Sherrington* y otros, 2000	11 *items*	Apenas fue mencionado que fue validado, por eso el tipo de validez no fue relatada	Kappa varió de k=0,61 para 0,88. CCI varió de 0,39-0,91
Bizzini	*Bizzini* y otros, 2003	4 criterios principales con 14 criterios específicos	Validez aparente y de contenido	CCI=0,64-0,99 dentro de los 4 principales criterios Para el puntaje total, CCI fue de 0,97
Chalmers	*Chalmers* y otros, 1981	32 *items* : Formulario 1 - Material básico descriptivo (9 ítems); Formulario 2 - Estudio del protocolo (14 ítems); Formulario 3 – Análisis estadística (9 ítems); Formulario 4 - Presentación de los resultados (4 *items*)	Validez aparente y de contenido. Autores afirman que esa escala exige una validez adicional	CCI varió de 0,66 a 0,92. *Test-Retest*: ICC 0,81
Reisch	*Reisch* y otros, 1989	34 *items* en 13 dominios	Validez aparente. Validez concurrente con Detsky, Imperiale, Jadad y Van Tulder fue de 0,81, 0,78, 0,64 y 0,77 respectivamente	CCI: 0,51
Yates	*Yates* y otros, 2005	8 ítems e 26 sub- *items*	Validez Aparente, Contenido y Discriminatorio	CCI=0,91 para la escala completa, kappa= 0,07-0,74 para cada *items* separadamente
Detsky	*Detsky* y otros, 1992	5 *items* principales	Validez aparente, validez concurrente con Reisch, Imperiale, Jadad y Van Tulder fue de 0,79, 0,78, 0,81 y 0,89 respectivamente	Correlación de Spearman varió de 0,85 para 0,96. CCI=0,92 (95 % IC=0,81-0,98), CCI=0,80
Sindhu	*Sindhu, Carpenter* y *Seers*, 1997	53 *items* con 15 dimensiones	Validez aparente, de contenido y validez de criterio en comparación con Chalmers (coeficientes de correlación fueron r_p=0,94, r_s=0,90 e r_i= 0,35)	No fue relatado Alfa de Cronbach. CCI: r_p=0,985, r_s=0,90 e r_i=0,93. Media de errores de 10 %. Evaluada solamente con dos evaluadores
Oxford Escala de Validez del Dolor	*Smith* y otros, 2000	5 *items* principales. *item* 5 subdividida en 4 *items*	Validez aparente	No relatado
Arrivé	*Arrivé* y otros, 2000	15 *items*	Validez aparente	K=0,8-1,0 para 11 *items* y k=0,74-0,78 para cuatro *items* . Para el puntaje total, r=0,91
New Castle Otawa	*Wells* y otros, 2000	8 *items* , divididos en tres dimensiones, incluyendo comparación y selección, y dependiendo del tipo de estudio, estudios de corte o estudios caso-control	Validez aparente y de contenido	No relatado

ESCALAS

- *Escala de Jadad*:[12] fue originalmente desarrollada y validada para evaluar de forma independiente la calidad de ECAs sobre el dolor, pero ha sido utilizada para otros propósitos, incluso como "padrón oro". Presenta puntuación de calidad de cinco puntos, con dos puntos adicionales para métodos apropiados de aleatorización y sigilo de colocación, que varía de 0 (débil) a 5 (bueno).[12] El primer *item* trata de la forma de aleatorización de los pacientes; el segundo, del uso del duplo-ciego; y el tercero de la pérdida de individuos.[12] Esta escala presentó evidencia de validez concurrente y demostró fuerte correlación con diversas escalas. El coeficiente de correlación interclase (CCI) varió de satisfactorio a excelente y Kappa reveló fiabilidad de débil a excelente.

Cuadro 2. Sinopsis de las características de las listas de evaluación de la calidad metodológica y de la información

Lista	Autores/Año	N° de *items*	Validez	Confianza
Delphi	Verhagen y otros. 1998[18]	9 *items*	Validez Aparente y de Contenido. Validez Concurrente con Escala de Jadad fue Spearman r=0,63	kappa 0,67, 0.76, 0,84, 0,70, 0,85. CCI=0,88
Maastricht Amsterdam	Van Tulder y otros. 1997[24]	19 *items*	Validez Aparente y Validez de Contenido	k=0,29, k=0,74, k=0,80, k=0,64, k=0,72, k=0,62, k=0,62
MOOSE	Stroup y otros. 2000[39]	35 *items* dicotómicos y divididos en seis secciones	No relatado	No relatado
Andrew Modificada	Andrew y otros. 1990[40]	11 *items*	Validez Aparente	Confiabilidad e inter-observador: r= 0,32
Downs y Black	Downs y Black 1998[41]	5 subescalas (Registro, validez externa, errores, confusión y poder), divididos en 27 *items*	Criterio de validez: El índice de calidad altamente correlacionada con el puntaje de Padrones de Registros Grupo de Ensayos (SRTG) (r =0,90)	A consistencia interna del índice de calidad: Kuder-Richardson (KR20)=0,89 Teste-reteste para el índice de calidad fue de r=0,88. O CCI de calidad fue de r=0,75
Nguyen	Nguyen y otros. 1999[45]	100 *items* divididos en validez interna y externa	Validez Aparente	No relatado
Declaración CONSORT	Huwiler-Muntener y otros. 2002[46]	25 *items*	Validez Aparente y de contenido	94% de concordancia entre os revisores
Criterios Cochrane	Colaboración Cochrane 2005-2007[11]	Evalúa la forma de colocación de los pacientes	No relatado	No relatado
Declaración STROBE	Vandenbrouc ke y otros. 2007[51]	22 *items*. Dieciocho *items* comunes a los estudios de corte, caso-control y estudios transversales y cuatro son específicos para cada uno de los tres estudios	No relatado	No relatado
AMSTAR	Shea y otros. 2007[54]	11 *items*	Validez Aparente, de Contenido y de Constructo. CCI fue de 0,66 para AMSTAR (95 %; CI: 0.28, 0.84) contra OQAQ y 0,83 (IC 95%: 0,64, 0,92) contra el instrumento Sacks	El CCI para el puntaje total fue excelente 0,84
PRISMA	Liberati y otros. 2009[55]	27 *items* y un diagrama de flujo de cuatro fases	No relatado	No relatado

- *Escala de Maastricht*:[13] evalúa metodológicamente la calidad de un ensayo clínico y tiene una función educativa en cuanto a la concepción y publicación. Consiste en 15 *ítems* basados en criterios metodológicos de evaluación de la calidad, que son divididos en 47 sub-*items* y totalizan 100 puntos en tres dimensiones de la calidad de un ensayo clínico: validez interna, validez externa y método estadístico. La escla incluye cuatro opciones de respuesta, y los valores son atribuidos a los *items* que reflejan la importancia relativa.[14] La validez concurrente con la Lista Delphi y la escala de Jadad presentó una fuerte correlación y el CCI fue excelente.

- *Single-Case Experimental Design Scale* (SCED): busca evaluar la calidad metodológica de los estudios de caso.[15] La *Yates* busca medir la calidad de los ECAs de intervenciones psicológicas en el tratamiento del dolor crónico y su uso ha sido muy limitado.[16,17] La *SCED* fue construida incluyendo 11 *ítems*, de los cuales 10 son utilizados para evaluar la calidad metodológica y la utilización de análisis estadístico.[18] La *SCED* muestra un excelente nivel de fiabilidad entre los evaluadores cuando se utiliza la puntuación total. Para todos los *items*, la fiabilidad varió de satisfactorio a excelente por dos evaluadores.

- *Escala de Van Tulder*:[19,20] analiza las amenazas a la validez de los ECAs partir de los elementos de adecuación del método aleatorio, ocultamiento de colocación del tratamiento, poca visión y análisis por intención del tratamiento. Trata de una escala de 11 puntos que analiza las amenazas a la validez e incluye elementos de adecuación metodológica de la investigación.[21,22] Esta lista apenas incluye los criterios de validez interna.[23] La validez aparente y la de contenido fueron analizadas. Esta escala presentó una fuerte correlación en la validez concurrente con otras escalas y en la fiabilidad interobservadores, y mostró fiabilidad, que varió entre débil a moderada.

- *Escala PEDro*:[8,24] fue desarrollada para ser empleada en estudios experimentales. Ofrece una importante fuente de información para apoyar la práctica basada en evidencias clínicas. Esta escala evalúa la validez interna y presentación del análisis estadístico de los estudios. Presenta 10 *items* sobre la validez interna y presentación del análisis estadístico. La presencia de indicadores de la calidad de las evidencias presentadas se asigna 1 punto y no 0 puntos.[25] Esta escala fue validada, aunque todavía no se identificó de cuál tipo. La fiabilidad presentó una variación del Kappa entre buena a excelente y el CCI de malo a excelente.

- *Escala Bizzini*:[26] busca evaluar la calidad de los ECAs sobre el síndrome del dolor femoropatelar a partir de cuatro criterios principales (población, intervenciones, tamaño del efecto, y de la presentación y análisis de datos) y 14 criterios específicos. Fueron atribuidos 25 puntos a cada uno de los cuatro criterios principales para un total de 100 puntos, y un máximo de 5 a 10 puntos para los criterios específicos. Todos los criterios varian de 0 a 5 o de 0 a 10 puntos, con 0 para una descripción inadecuada y el número máximo de puntos para una descripción detallada y apropiada.[26] La validez aparente y la de contenido fueron analizadas y el CCI varió de satisfactorio a excelente.

- *Escala Chalmers*,[27] evalúa la calidad por medio de 32 *ítems*. El puntaje evalúa dos dimensiones de la calidad (generalización interna y externa de validez) con puntuaciones máximas de 88 puntos.[27-29] Fue analizada la validez aparente y de contenido, y por eso esta escala exige una validación adicional. En relación con el CCI, hubo una variación satisfactoria a excelente, y en el *test-retest* fue excelente.

- *Escala Reisch*:[27,30] evalúa la calidad de ECAs sobre el uso de la aspirina en enfermedades coronarias. Recientemente fue adaptada y presentó fiabilidad para estudios sobre tratamientos farmacológicos para la osteoporosis.[29] Es constituida por 34 *items* divididos en 13 dominios (objetivos, proyecto experimental, determinación de la muestra, aleatorización y estratificación, descripción y aptitud de los participantes, uso de comparación con el grupo de control, procedimientos para el tratamiento o de gestión, ocultamiento, reducción de participantes, análisis y evaluación de los participantes en tratamiento, presentación y análisis de los datos, y las recomendaciones y conclusiones). La validez aparente y concurrente presentaron una fuerte correlación con otras escalas. El CCI fue considerado satisfactorio.

- *Escala Yates*:[16,17] busca medir la calidad de los ECAs de intervenciones psicológicas en el tratamiento del dolor crónico y su uso ha sido muy limitado. Es constituida por 8 *items* y 26 sub-*ítems* y su uso ha sido limitado, ya que fue citado solamente una vez por el mismo grupo de autores.[16] Para demostrar la validez de la escala, fueron analizadas la validez aparente, de contenido y de discriminación. En esta escala el coeficiente Kappa de cada uno de los *items* analizados varió de débil a satisfactorio y el CCI fue considerado excelente.

- *Escala Detsky*: evalúa la calidad de ensayos clínicos de soporte nutricional parenteral para pacientes sometidos a cirugía de gran porte.[23,31] Consiste en 13 variables con cinco *items* principales, que alcanza una puntuación máxima de 14 puntos. En la validez concurrente realizada con las escalas de Reisch, Jadad y Van Tulder presentó una fuerte correlación. La correlación de Spearman varió de fuerte a excelente.

- *Escala Sindhu*:[27] es una herramienta de evaluación de la calidad metodológica de los ECAs primarios a ser incluidos en un metaanálisis. Consta de 53 *items* subdivididos en 15 dimensiones sobre la calidad metodológica de ECAs utilizados en metaanálisis.[32] En el *test* de fiabilidad, el CCI demostró fuerte correlación; por eso fue evaluado solamente con dos evaluadores. La Sindhu presentó altos coeficientes de correlación de la validez aparente, de contenido y de criterio en comparación con la escala de Chalmers, pero todavía necesita de más pruebas.

- *Escala de Validez del Dolor Oxford (OPV)*: fue constituida con la finalidad de medir la validez de los resultados de ECAs para permitir la clasificación de los resultados de ensayos de acuerdo con la validez de las evaluaciones.[13] La OPV está compuesta por cinco *items* principales, y el último *item* es dividido en cuatro sub- *items* cualitativos.[33] Solamente fue analizada la validez aparente y la fiabilidad no fue relatada.

- *Escala Arrivé*: fue construida con el objetivo de evaluar la calidad metodológica de las investigaciones clínicas que utilizan exámenes radiológicos.[34] Evalúa la calidad metodológica por medio de 15 *items* relacionados con el diseño de estudio y con las características de la población, más allá de una descripción del análisis de la imagen.[34] La fiabilidad fue medida entre dos observadores, y presentó buena a alta concordancia. El CCI mostró una alta concordancia. Solamente fue analizada la validez aparente.

- *Escala Newcastle-Otawa (NOS)*: fue desarrollada para evaluar la calidad de estudios no aleatorizados buscando incorporar las evaluaciones de calidad en la interpretación de metaanálisis de los resultados obtenidos.[35,36] La NOS evalúa la calidad a partir del contenido, diseño y facilidad de uso en la interpretación del metaanálisis. Está compuesta por ocho *ítems*, divididos en tres dimensiones (comparación, selección, tipo de estudio) de investigaciones de corte, transversales o caso-control.[35,36] La validez aparente y de contenido fue establecida con base en una revisión crítica de los *ítems* por especialistas en el área.[36] La fiabilidad no fue relatada, pero ya fue utilizada recientemente en el metaanálisis que observó que la escala demuestra ser confiable y válida.

LISTAS

- *Lista Delphi*: representó el primer paso en dirección a un padrón en la evaluación de la calidad de ECAs,[18] pero no ha sido utilizado correctamente, ya que debe ser utilizada juntamente con otros instrumentos de evaluación de la calidad metodológica.[18] Esta lista evalúa los ECAs por medio de ocho preguntas sobre el método de aleatorización utilizado, realización del ocultamiento de la colocación, enmascaramiento del evaluador, del terapeuta y del paciente, y del análisis estadístico.[18,37] Esta lista presentó mayor validez en comparación con otras escalas, pero aún requiere establecer la consistencia interna y la validez. La fiabilidad interobservador varió de buena a excelente y presentó fuerte CCI.

- *Lista de Maastricht Amsterdam (MAL)*: es recomendada para revisiones sistemáticas que investigan problemas de la columna, específicamente del dolor lumbar.[14] La MAL es compuesta por 19 *items* que consisten en la evaluación de la validez interna, de los criterios descriptivos y de los aspectos estadísticos adoptados. La puntuación numérica general adoptada es de 0 a 19 puntos.[14,33,38] La validez aparente y de contenido fueron analizados y la fiabilidad inter-observador presentada varió de débil a buena.

- *Registro MOOSE*: es una lista de verificación detallada para relatar el metaanálisis de los estudios de observación en epidemiologia.[39] La lista MOOSE está organizada en 35 *items* dicotómicos y divididos en seis secciones que abordan informaciones sobre el resumen, la estrategia de la investigación, los métodos, los resultados, la discusión y las conclusiones.[39] La validez y la fiabilidad de esta lista no fueron identificadas en los estudios analizados.

- *Lista Andrew*: evalúa la calidad de estudios de tipo ensayo clínico que se utilizan medios de contraste de rayos-X.[40] La Andrew[40] fue criada en la decada de 1980 y modificada en la década de 1990 por los mismos autores y es compuesta por 11 *items* que buscan la evaluación de calidad de estudios de tipo ensayo clínico. La Lista modificada presentó una fiabilidad inter-observador débil.

- *Lista Downs y Black*: es utilizada para estudios aleatorios y no aleatorios,[41] y fue recientemente revisada para la utilización en la evaluación de la calidad de los estudios epidemiológicos,[42] pero todavía no es aplicable para estudios de prevalencia.[43] Está constituida por 27 *items* con cinco subescalas (registro, validez externa, errores, confusión, poder) y se puede utilizar en diversos tipos de estudios.[41] El índice de calidad fue altamente correlacionado con los puntajes de Padrones de Registros de Grupos de Ensayos. La consistencia interna fue considerada adecuada, así como lo CCI. Esta lista fue recientemente ampliada por dos nuevos criterios y validada para uso en estudios epidemiológicos.[43]

- *Lista Nguyen*: fue desarrollada para evaluar la calidad metodológica de los estudios.[44,45] La Nguyen[45] fue desarrollada con atribución para cada uno de los 18 ítems de una puntuación numérica de acuerdo con las orientaciones de acompañamiento, pero actualmente no es recomendado su uso. Fue analizada la validez aparente, y la confiabilidad no fue relatada. Con todo eso, algunos de los *items* de la lista deben ser evaluados de forma subjetiva, ya que no son directamente relacionados con la calidad del estudio.[44]

- *Declaración CONSORT*: fue desarrollada por un grupo de científicos y editores para evaluación crítica e interpretación de los ECAs,[26] en busca de evaluaciones más precisas y reproducibles.[21,46-48] Fue publicada la primera vez en 1996 y actualizada en el año 2001. En enero de 2007, la Declaración CONSORT fue posteriormente modificada para revista y fue publicada como la Declaración CONSORT 2010.[47] La CONSORT presenta 25 *items*.[49] Las preguntas son respondidas con sí (1 punto) o no (0 puntos), para un máximo de 25 puntos. La CONSORT presentó 94 % de concordancia entre los revisores.[21]

- *Criterios de la Colaboración COCHRANE*:[11] sirven para evaluar la condición de asignación del paciente y clasificar los estudios en adecuado, dudoso, inadecuado y no realizado. Para los criterios de la COCHRANE la aleatorización adecuada es necesaria para generar una comparación sin errores entre los grupos.[50] La validez y la fiabilidad no fueron relatadas.

- *Declaración STROBE:* busca la calidad de la información de estudios de observación con enfoque sobre prevalencia (corte, caso-control, transversales),[51] y sirve de apoyo para editores y revisores.[52] Está constituida por 22 *ítems* sobre el título de artículos, resumen, introducción, métodos, resultados, secciones de discusión y otras informaciones. Un total de 18 *items* son comunes a los tres diseños; en cuanto a los otros *items* son específicos del diseño. Para algunos *items*, las informaciones deben ser dadas separadamente para casos y controles en estudios caso-control, o grupos expuestos y no expuestos en el estudio de corte y estudios transversales.[52]

- AMSTAR: es un instrumento que busca la evaluación de la calidad metodológica de revisiones sistemáticas de los ECAs.[16,32] Está compuesto por 11 *ítems* de medidas confiables y válidas para la evaluación de la calidad metodológica de revisiones sistemáticas sobre los ECAs.[53,54] La validez aparente, de contenido y de constructo fueron testadas. El CCI fue testado en relación con otras escalas. En la fiabilidad, el CCI para el puntaje total fue excelente. Sin embargo, todavía son necesarios estudios adicionales con foco en la reproducibilidad y validez de constructo.[54]

- *PRISMA*: es una lista de verificación que tiene como objetivo buscar la transparencia de las informaciones de revisiones sistemáticas importantes al proceso de calificación científica de estos estudios. Posee 27 *items* y un diagrama de flujo de cuatro fases que incluye *items* considerados esenciales para la comunicación transparente de una revisión sistemática.[55] La validez y la fiabilidad no fueron relatadas.

Sinópticamente se puede afirmar que las escalas y listas de verificación actualmente presentan una gran amplitud de aplicaciones, pero también muchas limitaciones, ya que pueden ser utilizadas en la evaluación de estudios de diferentes tipos, para diferentes poblaciones y sobre diferentes enfoques en salud, como la Jadad, Maastricht, Van Tulder, Bizzini, Chalmers, Yates, Detsky, Sindhu, OPV, Delphi, Andrew Modificado, Downs y Black, CONSORT, Criterios Cochrane y AMSTAR en los ECAs; la Reisch, Detsky, NOS, Andrew Modificado y Downs y Black en los ensayos controlados no aleatorizados; la MAL, Detsky, Sindhu, PRISMA, Critérios Cochrane y AMSTAR en las revisiones sistemáticas; la Sindhu, NOS, MOOSE y Criterios Cochrane en las meta-analisys; la PEDro y los Criterios Cochrane en los estudios experimentales; la MOOSE, Downs y Black y STROBE en los estudios epidemiológicos; la SCED y la PRISMA en las investigaciones clínicas; la Arrivé y la Andrew Modificado en investigaciones clínicas especificas. Se puede verificar también que algunas escalas y listas ya no se recomiendan, como la Nguyen, y deben ser utilizadas conjuntamente con otras escalas y listas, como la Delphi.

Se puede decir también que sinópticamente las escalas y listas de verificaciones identificadas presentan una gran variedad de índices de validez y fiabilidad. Las escalas y listas que presentan validez adecuada son: Jadad, Delphi, Maastricht, MAL, Van Tulder, Bizzini, Chalmers, Reisch, SCED, Andrew Modificado, Yates, Detsky, CONSORT, Sindhu, Downs y Black, Nguyen, OPV, Arrivé, Criterios Cochrane, AMSTAR y NOS. Las escalas y listas que presentaron fiabilidad adecuada son: Jadad, Delphi, Maastricht, MAL, Van Tulder, PEDro, Bizzini, Chalmers, Reisch, SCED, Andrew Modificado, Yates, Detsky, CONSORT, Sindhu, Downs y Black, Arrivé, Criterios Cochrane, AMSTAR, PRISMA y NOS. Algunas escalas y listas o no fueron realizadas o no fue informada su validez (PEDro, STROBE, PRISMA) y confiabilidad (MOOSE, Nguyen, OPV, STROBE) de los procesos, o también no presentaron índices adecuados.

DISCUSIÓN

Considerando los resultados presentados, se constató la aparición de 14 escalas (incluyendo sus modificaciones) y 11 listas. Con todo esto, se puede verificar que apenas 14 escalas (Jadad, Maastricht, PEDro, Van Tulder, Bizzini, Chalmers, Reisch, SCED, Andrew Modificada, Yates, Detsky, Sindhu, Downs y Black y Arrivé) y 4 listas (Delphi, MAL, CONSORT y AMSTAR) presentaron propiedades psicométricas (validez y fiabilidad) debidamente investigadas. Se puede identificar que la escala de Jadad presentó la mejor evidencia de validez y fiabilidad, porque fue testada en diferentes contextos, y —en conjunto con la Lista Delphi— se tienen pruebas de mayor validez en comparación con las otras escalas y listas (MAL, Van Tulder, PEDro y Bizzini). En tanto, la Lista Delphi carece de consistencia interna y validez de constructo. Estas propiedades psicométricas son relevantes porque indican que la construcción, en este caso la calidad metodológica, está totalmente representada por los *ítems* de la escala (consistencia interna), y que los puntajes de una escala deben ser adecuados con base en hipótesis pre-definidas.[5,56] Las listas MOOSE, Criterios Cochrane, STROBE y PRISMA no presentaron las propiedades psicométricas, y las escalas Nguyen, Oxford y NOS solo relataron la validez.

Específicamente en el caso de la escala de Jadad, se puede averiguar que aunque haya sido desarrollada y validada para evaluar la calidad de los estudios realizados sobre el dolor, también ha sido utilizada extensivamente en otras áreas clínicas.[22] Actualmente, innumerables ensayos clínicos incluyen los *items* de la escala de Jadad en su metodología a fin de realizar un estudio con buena calidad metodológica. En este sentido, *Herbison* y otros[57] concluyeron que la escala de Jadad puede no ser sensible o suficiente para distinguir entre diferentes niveles de calidad. Por lo tanto, la utilización de la escala de Jadad y su validez debe ser reevaluada para diferentes áreas de investigación. En el caso de la lista Delphi, se puede verificar que a pesar de haber sido construida específicamente para la evaluación de la calidad de los ECAs,[18] ha sido utilizada en diversas otras áreas. Uno de los factores que contribuyen para esta ampliación de su uso es en relación con el hecho de que las propiedades psicométricas de la lista Delphi indican que la calidad metodológica está plenamente representada por los *items* de la escala (consistencia interna), y que los puntajes de una escala son basados en hipótesis predefinidas,[16] lo que llevó fuertemente a sustituir la escala de Maastricht que era ampliamente usada (MAL, Van Tulder, PEDro).[40] La escala Maastricht modificada y desarrollada a partir de una escala valida y confiable, no puede ser considerada valida y confiable hasta que sea testada. De acuerdo con *Streiner* y *Norman,*[56] las modificaciones de las escalas existentes, muchas veces exigen nuevos estudios de validez. Eso significa que las propiedades psicométricas de la escala modificada tienen que ser evaluadas para asegurar que la nueva escala pueda realmente identificar una buena o mala calidad metodológica.

Generalmente, se puede verificar que las escalas y listas existentes proveen recursos metodológicos para los lectores, autores, revisores y editores científicos a partir de diferentes propuestas y objetivos, y son ellas mismas un elemento de constante revisión, como la escala PEDro, una modificación de la lista Delphi, que ofrece una forma más abrazadora de medida de la calidad metodológica de la literatura de rehabilitación después del accidente vascular cerebral en comparación con la escala de Jadad,[58] la lista de *Downs* y *Black*, que fue revisada para la evaluación de la calidad de la población con base en estudios epidemiológicos,[42] pero todavía no es aplicable para estudios de prevalencia,[43] e incluso la NOS, que a pesar de estar en proceso de evaluación de la validez de la escala, pero ya con indicativos de ser confiable y valido,[25] posee *ítems* problemáticos, con falta de adecuación del análisis, falta de información relacionada a la fiabilidad y validez.[29]

La producción científica de diversos países ha crecido considerablemente en las últimas décadas.[59,60] En este sentido, la evaluación de la calidad de los estudios se torna esencial por su transparencia, visibilidad, rigor e impacto de la producción y publicación científica. Editores, revisores e investigadores deben cada vez más tener conocimiento de las herramientas de la evaluación de la calidad metodológica y de la información, porque los padrones para la publicación de los artículos se vienen tornando cada vez más exigentes y estrictos. Con todo esto, se debe destacar la crítica necesaria a estos procesos de calificación de la producción científica, porque se debe considerar que la mayoría de las escalas y listas revisadas no alcanzan los padrones metodológicos adecuados, y exige la inclusión de *ítems* importantes que deben ser evaluados en cuanto al tipo de estudio, aplicación, capacidad psicométrica, principalmente de validez y fiabilidad, entre otros. En este sentido, investigadores, editores y analistas deben tener especial cuidado y atención al utilizar determinadas escalas para evaluar la calidad de los estudios científicos, porque las limitaciones y las informaciones presentadas deben ser interpretadas con cautela. De acuerdo con *Serra*,[7] la necesidad de publicar resultados de forma clara y transparente influye positivamente sobre la formación de nuevos conocimientos, y consecuentemente aumenta la confianza en las conclusiones cuando el estudio es realizado con rigor metodológico adecuado. Por lo tanto, a pesar de los referidos cuidados, la propagación de la utilización de las herramientas de la evaluación metodológica y de información califica y legitima cada vez más la producción científica, principalmente en el área de la salud.

Se concluye que, considerando el aumento de las investigaciones vinculadas a la ciencias de la salud y áreas afines, se torna esencial establecer una evaluación de la calidad metodológica de los estudios realizados ya que este procedimiento puede contribuir para que se eviten publicaciones inconsistentes, y para mejorar el proceso de selección de la literatura en términos de su validez, relevancia y aplicabilidad clínica. De este modo, con la intención de asegurar el rigor científico de las investigaciones con base en los resultados de esta revisión sistemática, se puede asegurar que muchas escalas están siendo producidas para evaluar la calidad metodológica de artículos científicos y listas de verificación de calidad de información en investigaciones en el área de la salud. Se puede destacar, que las escalas y las listas difieren entre sí en relación al número de *items*, validez, fiabilidad y parámetros de puntuación, y que diversas de éstas presentan propiedades psicométricas validas y confiables, más allá de que estas escalas y listas son aplicables a estudios empíricos, principalmente a los ensayos clínicos aleatorizados, y a estudios revisionales, principalmente los metanalíticos, tanto en la búsqueda de la calidad metodológica, cuanto de la calidad de la información. Por lo tanto, cabe destacar que la utilización de evaluaciones metodológicas y de información es mayor, influenciando positivamente en el aumento de la producción científica, en especial en el área de la salud y en áreas afines.

REFERENCIAS BIBLIOGRÁFICAS

1. Verhagen AP, de Vet HCW, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol. 2001;54(7). p. 651-4.

2. Emerson JD, Burdick E, Mosteller F, Chalmers TC. An empirical study of the possible relation of treayment differences to quality scores controlled randomised clinical trials. Contr Clin Trials. 1990;11:339-52.

3. Juni P, Altman DG, Egger M. Systematic reviews in health care - Assessing the quality of controlled clinical trials. Brit Med J. 2001;323(7303). pp. 42-6.

4. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials - an annotated-bibliography of scales and checklists. Contr Clin Trials. 1995;16(1):62-73.

5. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. Oxford: 2003.

6. Greenfield M, Rosenberg AL, O'Reilly M, Shanks AM, Sliwinski MJ, Nauss MD. The quality of randomized controlled trials in major anesthesiology journals. Anesth Analg. 2005;100(6):1759-64.

7. Serra LC. Las buenas prácticas de publicación, su evolución y el impacto esperado en salud pública. Rev Cubana Sal Públ. 2012;38:725-33.

8. Taylor CF, Dawn F, Sansone SA, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotech. 2008;26:889-96.

9. Munhoz Junior E. Requisitos uniformes para manuscritos submetidos a periódicos biomédicos: escrevendo e editando para publicações biomédicas. Epidemiol Serv Saúde. 2006;15:7-34.

10. Higgins JPT, Green S. Manual Cochrane de revisiones sistemáticas de intervenciones. The Cochrane Collaboration; 2011.

11. Higgins JPT. Assessing risk of bias in included studies. The Cochrane Collaboration; 2005.

12. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Contr ClinTrials. 1996;17(1):1-12.

13. de Vet HC, de Bie RA, van der Heijden GJ, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. Physioth. 1997;83(6):284-89.

14. Van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM. Method guidelines for systematic reviews in the Cochrane Collaboration back review group for spinal disorders. Spine. 1997;22(20). p. 2323-30.

15. Tate AR, Gennings C, Hoffman RA, Strittmatter AP, Retchin SM. Effects of bone-conducted music on swimming performance. J Strength Cond Res. 2012;26(4):982-8.

16. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: A systematic review. Physioth. 2008;88(2):156-75.

17. Yates SL, Morley S, Eccleston C, Williams ACD. A scale for rating the quality of psychological trials for pain. Pain. 2005;117(3):314-25.

18. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, et al. The delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998;51(12). p. 1235-41.

19. Sung L, Nathan PC, Lange B, Beyene J, Buchanan GR. Prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor decrease febrile neutropenia after chemotherapy in children with cancer: A metaanalysis of randomized controlled trials. J Clin Oncol. 2004;22(16). p. 3350-6.

20. Dekker A, Bulley S, Beyene J, Dupuis LL, Doyle JJ, Sung L. Meta-analysis of randomized controlled trials of prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor after autologous and allogeneic stem cell transplantation. J Clin Oncol. 2006;24(33). p. 5207-15.

21. Warschkow R, Tarantino I, Jensen K, Beutner U, Clerici T, Schmied BM, et al. Bilateral Superficial Cervical Plexus Block in Combination with General Anesthesia Has a Low Efficacy in Thyroid Surgery: A Metaanalysis of Randomized Controlled Trials. Thyroid. 2012;22(1):44-52.

22. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, et al. Assessing the quality of randomized trials: Reliability of the Jadad scale. Contr Clin Trials. 1999;20(5):448-52.

23. van Tulder M, Furlan A, Bombardier C, Bouter L, Editorial Board Cochrane C. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. Spine. 2003;28(12):1290-9.

24. Sherrington C, Herbert RD, Maher CG, Moseley AM. PEDro. A database of randomized trials and systematic reviews in physiotherapy. ManTher. 2000;5(4):223-6.

25. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther. 2003;83(8):713-21.

26. Bizzini M, Childs JD, Piva SR. Systematic review of the quality of randomized controlled trials for patellofemoral pain syndrome. Journal of Orthopaedic & Sports Phys Ther. 2003;33(1):4-20.

27. Chalmers TC, Jr Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. In: Smith HJR BB, editor. Contr Clin Trials.1981;2(1):31-49.

28. Sculier JP, Berghmans T, Castaigne C, Luce S, Sotiriou C, Vermylen P, et al. Maintenance chemotherapy for small cell lung cancer: a critical review of the literature. Lung Canc. 1998;19(2):141-51.

29. Berard A, Andreu N, Tetrault JP, Niyonsenga T, Myhal D. Reliability of Chalmers scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. An Epidemiol. 2000;10(8):498-503.

30. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatr. 1989;84(5):815-27.

31. Detsky AS, Naylor CD, Orourke K, McGeer AJ, Labbe KA. Incorporating variations in the quality of individual randomized trials into metaanalysis. J Clin Epidemiol. 1992;45(3):255-65.

32. Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. J Adv Nurs. 1997;25(6):1262-8.

33. Smith LA, Oldman AD, McQuay HJ, Moore RA. Teasing apart quality and validity in systematic reviews: an example from acupuncture trials in chronic neck and back pain. Pain. 2000;86(1-2):119-32.

34. Arrive L, Renard R, Carrat F, Belkacem A, Dahan H, Le Hir P, et al. A scale of methodological quality for clinical studies of radiologic examinations. Radiology. 2000;217(1):69-74.

35. Aarts JWM, van den Haak P, Nelen WLDM, Tuil WS, Faber MJ, Kremer JAM. Patient-focused Internet interventions in reproductive medicine: a scoping review. Hum Reprod Upd. 2012;18(2):211-27.

36. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in metaanalyses. 2000.

37. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy and quality assessment: Interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. J Clin Epidemiol. 1998;51(4). p. 335-41.

38. Seferiadis A, Rosenfeld M, Gunnarsson R. A review of treatment interventions in whiplash-associated disorders. Europ Sp J. 2004;13(5):387-97.

39. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Metaanalysis of observational studies in epidemiology: a proposal for reporting. J Am Med Assoc. 2000;283(15). p. 2008-12.

40. Andrew E, Eide H, Fuglerud P, Hagen EK, Kristoffersen DT, Lambrechts M, et al. Publications on clinical trials with X-ray contrast media: differences in quality between journals and decades. Europ J Radiol. 1990;10(2):92-7.

41. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Comm Heal. 1998;52(6). p. 377-84.

42. Macfarlane TV, Glenny AM, Worthington HV. Systematic review of population-based epidemiological studies of oro-facial pain. J Dent. 2001;29(7):451-67.

43. Shamliyan T, Kane RL, Jansen S. Quality of Systematic Reviews of Observational Nontherapeutic Studies. Prev Chron Dis. 2010;7(6). p. 9-195.

44. Giannakopoulos NN, Rammelsberg P, Eberhard L, Schmitter M. A new instrument for assessing the quality of studies on prevalence. Clin Oral Invest. 2012;16(3). p. 781-8.

45. Nguyen QV, Bezemer PD, Habets L, Prahl-Andersen B. A systematic review of the relationship between overjet size and traumatic dental injuries. Europ J Orthod. 1999;21(5):503-15.

46. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. J Am Med Assoc. 2002;287(21):2801-4.

47. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2012;65(3). p. 239-46.

48. Moher D, Jones A, Lepage L, Grp C. Use of the CONSORT statement and quality of reports of randomized trials. A comparative before and after evaluation. J Am Med Assoc. 2001;285(15):1992-5.

49. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Springer Verlag. 2003;7(1). pp. 2-7.

50. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical-evidence of bias - dimensions of methodological quality associated with estimates of treatment effects in controlled trials. J Am Med Assoc. 1995;273(5):408-12.

51. von Elm E, Altman DG, Pocock SJ, Gotzsche PC, Vandenbroucke JP, Initiative S. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Brit Med J. 2007;335(7624):806-8.

52. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Explanation and Elaboration. Epidemiol. 2007;18(6):805-35.

53. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol. 2009;62(10):1013-20.

54. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. Bmc Med Res Methodol. 2007;7(1):10.

55. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and metaanalyses of studies that evaluate healthcare interventions: explanation and elaboration. Brit Med J. 2009;(6):1-27.

56. Streiner DL, Norman GR. Validity. Health measurement scales: A practical guide to their development and use: Oxford; 2004. p. 172-93.

57. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol. 2006;59(12):1249-56.

58. Bhogal SK, Teasell RW, Foley NC, Speechley MR. The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad Scale in stroke rehabilitation literature. J Clin Epidemiol. 2005;58(7):668-73.

59. Nunes ED. A review of research studies conducted on scientific production in collective health in Brazil. Scientometrics. 1999;44(2):157-67.

60. Glanzel W, Leta J, Thijs B. Science in Brazil. Part 1: A macro-level comparative study. Scientometrics. 2006;67(1):67-86.

Recibido: 25 de diciembre de 2012.
Aprobado: 26 de febrero de 2013.

MSc. *Franciele Cascaes da Silva*. Centro de Ciencias de la Salud y el Deporte de la Universidad del estado de Santa Catarina, Brasil. Correo electrónico: francascaes@yahoo.com.br