Rev. Chilena de Cirugía. Vol 63 - Nº 5, Octubre 2011; pág. 493-497

ARTÍCULO DE INVESTIGACIÓN

Diseño de una escala para evaluar calidad metodológica de estudios de pruebas diagnósticas. Estudio piloto*

Drs. MARÍA EUGENIA BURGOS D.[1,2], CARLOS MANTEROLA D.[1,2], ANTONIO SANHUEZA C.[2,3]

[1] Departamento de Cirugía y Traumatología.
[2] Programa de Doctorado en Ciencias Médicas. Facultad de Medicina.
[3] Departamento de Matemáticas y Estadística.
Universidad de La Frontera.
Temuco, Chile.

Abstract

Construction of a scale to assess methodological quality of diagnostic tests articles

Introduction: Despite the methodological quality (MQ) of scientific publications is a multidimensional concept difficult to understand, their evaluation is essential at the time of making decisions that support our clinical practice. However, in the field of diagnostic tests (DT), which is in a steady and rapid development, there are no valid and reliable instruments to assess MQ. **Aim:** To report the results of the generation of items and domains of a scale to determine MQ in studies of DT and to determine interobserver reliability of this scale. **Material and Methods:** Construction of a scale to assess MQ of DT articles and pilot study to determine interobserver reliability. Designed scale was applied to 20 DT studies randomly selected. Interobserver reliability of the scale and each of the domains that compose it was determined by applying intraclass correlation coefficient. **Results:** The created scale has 9 items grouped into three domains. The ICC observed was 1.0 for the domain 1, 0.90 for the domain 2 and 0.86 for the domain 3. The overall ICC was 0.96. **Conclusion:** A scale to determine MQ in DT studies was created and it interobserver reliability was determined with a significant level of agreement between observers.

Key words: "Epidemiologic Methods" [Mesh], methodological quality, "Reproducibility of Results" [Mesh], Reliability, scales.

Resumen

Introducción: A pesar que la calidad metodológica (CM) de las publicaciones científicas es un concepto multidimensional de difícil comprensión, su evaluación es fundamental para la toma de decisiones que apoyen nuestra práctica clínica. No obstante ello, en el ámbito de las pruebas diagnósticas (PD), que se encuentra en constante y rápido desarrollo, no existen instrumentos válidos y confiables que permitan evaluar CM. **Obje-**

*Recibido el 27 de octubre de 2010 y aceptado para publicación el 9 de marzo de 2011.

Parcialmente financiado por proyecto DI09-0060 de la Dirección de Investigación de la Universidad de La Frontera.

Correspondencia: Dra. María Eugenia Burgos D.
Casilla 54-D, Temuco, Chile. Fax: 56-45-325761
mburgos@ufro.cl

tivo: Reportar los resultados del proceso de generación de ítems y dominios de una escala para determinar CM en estudios de PD; y determinar la confiabilidad interobservador de esta escala. **Material y Método:** Construcción de una escala de CM de estudios de PD y estudio de confiabilidad interobservador. Se aplicó la escala diseñada a 20 artículos de PD seleccionados en forma aleatoria. Se determinó confiabilidad interobservador de la escala en general y de cada uno de los dominios que la componen mediante aplicación del coeficiente de correlación intraclase. **Resultados:** La escala generada quedó compuesta por 9 ítems agrupados en tres dominios. El CCI observado para el dominio uno fue de 1,0; para el dominio dos de 0,90; y para el dominio tres de 0,86. El CCI general de la escala fue de 0,96. **Conclusión:** Se generó una escala para medir CM en estudios de PD y se determinó confiabilidad interobservador de ella y los dominios que la componen. Se observó un nivel de acuerdo significativo entre los evaluadores.

Palabras clave: Metodología, calidad metodológica, confiabilidad, reproducibilidad, escalas.

Introducción

Evaluar la calidad metodológica (CM) de un artículo científico es un proceso complejo, esencialmente porque el constructo "CM" debe entenderse como un concepto multidimensional en el que es posible evaluar múltiples ítems; tales como la calidad del reporte, el tipo de diseño empleado, la metodología aplicada, el análisis utilizado, etc.[1,2].

En la actualidad, existen sistemas para calificar la calidad de las publicaciones, dentro de los cuales el más conocido es el de Sackett y cols[3,4]. Esta, es una escala alfanumérica que se basa únicamente en la evaluación del tipo de diseño de los estudios; sin embargo, continúa siendo una herramienta útil para establecer el nivel de evidencia que apoya nuestra forma de actuar y de tomar decisiones. Además, así como en otros ámbitos, existen para escenarios de diagnóstico algunos sistemas de chequeo para la elaboración de estudios y publicación de resultados, como las iniciativas STARD (Standards for Reporting of Diagnostic Accuracy), QUADAS (Quality Assessment of Diagnostic Accuracy Studies) y QAREL (Quality Appraisal of Reliability Studies)[6-9]. Sin embargo, ninguna de ellas fue diseñada para valorar CM, y por ende, tampoco han sido validadas para tal fin.

A nivel local, existe una línea de investigación en CM, cuyo objetivo es la generación de escalas válidas y confiables, que puedan ser aplicadas en distintos escenarios clínicos (tratamiento o procedimientos terapéuticos, pronóstico, diagnóstico, etc.). El primer producto dentro de esta línea de investigación fue una escala para valorar CM de estudios de tratamiento, la que ha permitido la realización de estudios bibliométricos y ponderar la evidencia para la conducción de revisiones sistemáticas y meta-análisis[10,11].

Para continuar con esta línea, y considerando que el área de investigación en pruebas diagnósticas (PD) ha sido una de las que ha tenido mayor y más rápido desarrollo en la medicina actual; nos planteamos la generación de una escala válida y confiable que permita valorar la CM de estudios de este tipo.

El objetivo de este estudio es reportar los resultados del proceso de generación de ítems y dominios de una escala para determinar CM en estudios de PD y determinar confiabilidad interobservador de esta escala.

Material y Método

Diseño

Estudio bietápico.

La primera etapa correspondió a la generación de la escala; etapa que se realizó a partir de la de una extensa revisión de la literatura con el objeto de identificar instrumentos ya existentes desde donde se pudiesen obtener potenciales ítems y dominios relevantes. Se utilizó entre otros instrumentos, la propuesta de niveles de evidencia del Centro de Medicina Basada en Evidencia de la Universidad de Oxford en su última versión (marzo de 2009) para la jerarquización de los tipos de diseños que se utilizan con mayor frecuencia en escenarios de diagnóstico. De este modo se generó el primer borrador de la escala, el que quedó constituido por 3 dominios (tipo de diseño, población estudiada y metodología empleada) y 9 ítems. Posteriormente, esta escala fue sometida a la evaluación por un grupo focal donde se realizaron las modificaciones consideradas pertinentes.

En la segunda etapa, se realizó un estudio piloto para determinar confiabilidad interobservador, aplicando la escala por dos investigadores independientes a artículos referentes a PD que fueron seleccionados en forma aleatoria. Los investigadores seleccionados tienen formación en epidemiología clínica y análisis crítico de la literatura biomédica; ambos al menos grado de Magíster y 5 años de experiencia en postgrado e investigación.

Metodología

Se estudiaron 20 artículos referentes a PD[12-31], obtenidos a partir de las bases de datos MEDLINE y SciELO. La estrategia de búsqueda consistió en la

utilización de los siguientes términos Mesh: "Diagnostic Tests"[Mesh], "Surgical Procedures, Operative"[Mesh] y "Diagnostic Test, Routine"[Mesh]; los que fueron combinados con el operador booleano "AND". La escala fue aplicada por dos investigadores en forma independiente.

Herramientas estadísticas

Después de realizar un análisis exploratorio de los datos, se aplicó estadística descriptiva con cálculo de medidas de tendencia central. Posteriormente, se realizó determinación de confiabilidad interobservador mediante coeficiente de correlación intraclase (CCI), de cada dominio por separado y de la escala en general.

Aspectos éticos

Se salvaguardó la confidencialidad de la información obtenida del estudio en relación a los autores y las revistas en las que habían sido publicados los artículos revisados.

Resultados

La escala generada quedó compuesta por 9 ítems agrupados en tres dominios; el primero, está relacionado con el tipo de diseño del estudio, ordenado de mayor a menor nivel de evidencia; el segundo, está relacionado con el tamaño de la población estudiada, ajustada según exista o no justificación del tamaño de la muestra empleado para la ejecución del estudio; y, el tercero, tiene relación con la descripción de la metodología utilizada en la conducción del estudio (planteamiento y claridad de los objetivos, mención del tipo de diseño empleado, aplicación de criterios de selección de la muestra, justificación de la muestra empleada, características de la población estudiada y lo apropiado del espectro de ésta, características del estándar de referencia en relación a la aplicación o no de un mismo estándar a la totalidad de los sujetos en estudio; y, características de la PD en estudio respecto de su descripción con fines de poder reproducir la experiencia por otros investigadores) (Figura 1).

	Puntuación
Diseño del estudio	
Revisión sistemática de estudios diagnósticos de alta calidad	**15**
Estudios de pruebas diagnósticas de alta calidad	**12**
Estudios de cohorte concurrente o prospectiva	**9**
Revisión sistemática de estudios diagnósticos de mediana calidad	**6**
Estudios de pruebas diagnósticas de mediana calidad	**4**
Estudios de cohorte histórica o retrospectiva	**3**
Estudios de casos y controles	**3**
Series de casos	**1**
Población estudiada x factor de justificación	
≥ 201	**7 ó 14**
151- 200	**6 ó 12**
101 – 150	**5 ó 10**
61 – 100	**4 u 8**
31 – 60	**3 ó 6**
≤ 30	**2 ó 4**
Descripción de la metodología empleada	
Objetivo	
• Se plantean objetivos claros y concretos	**3**
• Se plantean objetivos vagos	**2**
• No se plantean objetivos	**1**
Diseño	
• Se menciona el diseño empleado	**3**
• No se menciona el diseño empleado	**1**
Criterios de selección de la muestra	
• Se describen criterios de inclusión y de exclusión	**3**
• Se describen criterios de inclusión o de exclusión	**2**
• No se describen criterios de selección	**1**
Características de la población estudiada	
• Existe un espectro representativo del evento de interés en estudio	**3**
• Espectro de sujetos incompleto	**1**
Características del estándar de referencia aplicado	
• Se aplica el mismo estándar de referencia, independiente del resultado y a todos los sujetos en estudio	**3**
• Se aplica estándar de referencia de forma parcial	**2**
• No reporta utilización de estándar de referencia	**1**
Características de la prueba diagnóstica en estudio	
• Se describe la prueba en estudio con el detalle suficiente para permitir su replicación	**3**
• Se describe la prueba en estudio de forma parcial	**2**
• No se mencionan elementos de la prueba en estudio que permitan su replicación	**1**
Tamaño de la muestra	
• Justifica la muestra empleada	**3**
• No justifica la muestra empleada	**1**
Puntuación final	**10 a 50**

Figura 1. Escala preliminar con sus ítems y dominios.

El CCI observado para el dominio 1 fue de 1,0; para el dominio 2 de 0,90; y para el dominio 3 de 0,86. El CCI general de la escala fue de 0,96.

Discusión

Cada vez se hace más patente la necesidad de contar con instrumentos válidos y confiables que faciliten nuestro actuar clínico ya sea al momento de diseñar un estudio, de escribir un artículo, y de valorar la CM de una publicación ya sea como lector o revisor de una revista. Por esta razón existen algunas iniciativas que han abordado esta problemática y han publicado listas de chequeo, las que constantemente están siendo utilizadas en diversos ámbitos de la medicina[5-8,28,32-34]. Sin embargo, a pesar del valioso aporte de estas iniciativas, las cuales incluso han servido de base para la creación de nuestra escala, no ha existido un proceso de validación propiamente tal que les otorgue a estos procesos de medición un peso técnico y objetivo.

Es por ello que nos parece necesario continuar con esta línea de investigación, de modo de contar con escalas apropiadas para ser aplicadas a distintos escenarios clínicos, las que esperamos sean un aporte para investigadores, lectores, revisores y editores de revistas biomédicas; asumiendo que deben ser objeto de revisión y actualización periódica por parte de los usuarios.

Por otro lado, creemos que estas escalas deben ser lo suficientemente comprensibles y sencillas de aplicar, de modo tal de lograr una masificación de su uso que haga pensar que el tiempo invertido en el proceso de generación y validación de ellas no haya sido en vano.

Al realizar este estudio piloto, nos dimos cuenta en una primera aproximación, que la escala creada tiene un buen comportamiento al analizar la confiabilidad interobservador. En este caso, para establecer el grado de concordancia entre los observadores utilizamos el CCI aplicando la escala de Landis y Koch; con la cual un valor entre 0,8 y 1 se considera "casi perfecto". Esto significa que la escala que se ha presentado en este estudio tiene un nivel de reproducibilidad o confiabilidad de medición "casi perfecta" entre diferentes observadores; hecho que nos permite continuar con el proceso final de validación de esta escala.

Referencias

1. Manterola C. Respecto de la calidad metodológica de los artículos que se publican en las revistas biomédicas. Rev Chil Cir. 2005;57:449-50.
2. Manterola C, Busquets J, Pascual M, Grande L. ¿Cuál es la calidad metodológica de los artículos sobre procedimientos terapéuticos publicados en cirugía española? Cir Esp. 2006;79:95-100.
3. Sackett D. Rules of evidence and clinical recommendations on use of antithrombotic agents. Chest 1986;89:2S-3S.
4. Meakins J. Innovation in surgery: the rules of evidence. Am J Surg. 2002;183:399-405.
5. Begg C, Cho M, Eastwood S, Hortin H, Moher D, Olkin I, et al. Improving the quality of reports on randomized controlled trials. Recommendations of the CONSORT Study Group. Rev Esp Salud Publica 1998;72:5-11.
6. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasiou PP, Irving LM, et al. The Stard Statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138:W1-12.
7. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic realibility (QAREL). J Clin Epidemiol. 2010; 63:854-61.
8. Whiting P, Westwood M, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.
9. Cook C, Cleland J, Huijbregts P. Creation and Critique of Studies of Diagnostic Accuracy: Use of the STARD and QUADAS Methodological Quality Assessment Tools. 2007;15:93-102.
10. Manterola C, Pineda V, Vial M, Losada H, Muñoz S. Revisión sistemática de la literatura. Propuesta metodológica para su realización. Rev Chil Cir. 2003;55:204-8.
11. Pineda V, Manterola C, Vial M, Losada H. ¿Cuál es la calidad metodológica de los artículos referentes a terapia publicados en la Revista Chilena de Cirugía? Rev Chil Cir. 2005;57:500-7.
12. Contreras O, Burdiles A. Diagnóstico de lesiones óseas con biopsia percutánea guiada por imágenes. Rev Med Chile 2006;134:1283-7.
13. Einstein A, Henzlova M, Rajagopalan S. Estimating Risk of Cancer Associated With Radiation Exposure From 64-Slice Computed Tomography Coronary Angiography. JAMA 2007;298:317-23.
14. Fenton J, Rolnick S, Harris E, Barton M, Barlow W, Reisch L, et al. Specificity of Clinical Breast Examination in Community Practice. J Gen Int Med. 2007;22:332-7.
15. López J, Best A, Morales C. Diagnóstico de brucelosis bovina en leche por el Ring Test y ELISA en lecherías de la provincia de Ñuble (VIII Región). Arch Med Vet. 1998;30;133-8.
16. Melnikov A, Scholtens D, Wiley E, Khan S, Levenson V. Array-Based Multiplex Analysis of DNA Methylation in Breast Cancer Tissues. JMD 2008;10:93-101.
17. Norero E, Norero B, Huete A, Pimentel F, Cruz F, Ibáñez L, y cols. Rendimiento de la colangiografía por

resonancia magnética en el diagnóstico de coledocolitiasis. Rev Med Chile 2008;136:600-5.
18. Oyanedel R, García C, Villanueva E, Otero J, Solar A, Rojas R, et al. Estudio radiológico simple en el diagnóstico de condroblastoma epifisiario benigno. Correlación anátomo-radiológica. Rev Chil Radiol. 2007;13;185-90.
19. Reinholz M, Nibbe A, Jonart L, Kitzmann K, Suman V, Ingle J, et al. Evaluation of a Panel of Tumor Markers for Molecular Detection of Circulating Cancer Cells in Women with Suspected Breast Cancer. Clin Cancer Res. 2005;11:3722-32.
20. Tafra L, Lannin D, Swanson M, Van Eyk J, Verbanac K, Chua A, et al. Multicenter Trial of Sentinel Node Biopsy for Breast Cancer Using Both Technetium Sulfur Colloid and Isosulfan Blie Dye. Ann Surg. 2001;233:51-9.
21. Wittner B, Sgroi D, Ryan P, Bruinsma T, Glas A, Male A, et al. Analysis of the Mammaprint Breast Cancer Assay in a Predominantly Postmenopausal Cohort. Clin Cancer Res. 2008;14:2988-93.
22. Busel D, Pérez L, Arroyo A, Ortega D, Niedmann JP, Palavecino P, y cols. Colangioresonancia (CPRM) vs Ultrasonido (US) focalizado en pacientes con ictericia o sospecha de obstrucción de la vía biliar. Resultados preliminares. Rev Chil Radiol. 2003;9:173-81.
23. Coll C, Cifras J, Massardo T, Moya H. Cintigrafía ósea trifásica con Tc-99m MDP en el diagnóstico y manejo de infecciones osteoarticulares agudas en niños. Rev Chil Radiol. 2002;8:83-8.
24. Escalona A, Bellolio F, Dagnino B, Pérez G, Viviani P, Lazo D, y cols. Utilidad de la proteína C reactiva y recuento de leucocitos en sospecha de apendicitis aguda. Rev Chil Cir. 2006;58:122-6.
25. Feitosa D, da Silva M, Parada C. Accuracy of simple urine tests for diagnosis of urinary tract infections in low-risk pregnant women. Rev Latino-am Enfermagem. 2009;17:507-13.
26. Goodman L, Stein P, Matta F, Sostman HD, Wakefield T, Woodard P, et al. CT Venography and Compression Sonography are diagnostically equivalent: Data from PIOPED II. AJR 2007;189:1071-6.
27. Gudmundsson P, Shahgaldi K, Winter R, Dencker M, Kitlinski M, Thorsson O, et al. Quantitative detection of myocardial ischaemia by stress echocardiography; a comparison with SPECT. Cardiovascular Ultrasound 2009;7:28.
28. Kim D, Park A, Lee E, Choo H, Kim S, Lee S, et al. Ultrasound-Guided Fine-Needle aspiration biopsy of thyrod nodules smaller than 5 mm in the maximum diameter: Assessment of efficacy and pathological findings. Korean J Radiol. 2009;10:435-40.
29. Moghimi M, Ghodossi I, Rahimabadi AE, Sheikhvatan M. Accuracy of sentinel node biopsy in breast cancer patients with a high prevalence of axillary metastases. Scandinavian Journal of Surgery 2009;98:30-3.
30. Soler T, Isamitt D, Carrasco O. Rendimiento de la biopsia, cepillado y lavado bronquial por fibrobroncoscopia en el diagnóstico de cáncer pulmonar con lesiones visibles endoscópicamente. Rev Med Chile 2004;132:1198-203.
31. Targa-Stramandinoli RM, Moacir-Sassi L, Pedruzzi P, Ramos G, Oliveira B, Ogata D, et al. Accuracy, sensitivity and specificity of the fine needle aspiration biopsy in salivary gland tumours: A retrospective study. Med Oral Patol Oral Cir Bucal 2010;15:e32-37.
32. Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS One 2009;4:e7753.
33. Mac Pherson H, Altman DG, Hammrschlag R, Youping L, Taixiang W, White A, et al, STRICTA Revision Group. Revised Standards for Reporting Interventions in Clinical Trials of Acupunture (STRICTA): extending the CONSORT statement. PLoS Med 2010;7:e1000261.
34. Simel DL, Rennie D, Bossuyt PM. The STARD statement for reporting diagnostic accuracy studies: application to the history and physical examination. J Gen Intern Med. 2008;23:768-74.