Reumatol Clin. 2009;5(4):171–177

Reumatología Clínica

www.reumatologiaclinica.org

Revisión

Validación de cuestionarios

M[a] Jesús García de Yébenes Prous [a,*], Francisco Rodríguez Salvanés [b] y Loreto Carmona Ortells [a]

[a] *Unidad de Investigación, Fundación Española de Reumatología, Madrid, España*
[b] *Agencia Laín Entralgo, Madrid, España*

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 17 de julio de 2008
Aceptado el 16 de septiembre de 2008
On-line el 20 de mayo de 2009

Palabras clave:
Cuestionario
Escalas de evaluación
Validez
Fiabilidad

RESUMEN

El desarrollo de un cuestionario o instrumento de medición es un proceso laborioso y complejo y requiere la comprobación de su utilidad antes de su aplicación. Se presenta un trabajo metodológico sobre las características psicométricas de los instrumentos de evaluación y el análisis de sus principales características: la fiabilidad y la validez.

Validation of questionnaires

Keywords:
Questionnaire
Health indices
Validity
Reliability

ABSTRACT

The development of a questionnaire or a measuring instrument is a laborious and complex process and requires verification of its usefulness before implementation. We present a methodological work on the psychometric characteristics of assessment instruments and analysis of their main features, reliability and validity.

Introducción

En 1948, la Organización Mundial de la Salud definió «salud» como el completo estado de bienestar físico, mental y social. Desde entonces se han desarrollado numerosas investigaciones para traducir esta definición conceptual en métodos objetivos que, mediante cuestionarios u otros instrumentos, generen escalas e índices para facilitar la medición de las dimensiones del estado de salud. Junto con la entrevista, el cuestionario es la técnica más empleada en investigación. En este artículo se considerará que cuestionarios, escalas e instrumentos son sinónimos de un mismo concepto: técnica de recogida de datos.

La entrevista es una técnica de recogida de datos que requiere el conocimiento de técnicas de comunicación verbal, un guión estructurado y una finalidad específica. Es un instrumento excelente en investigación cualitativa. El cuestionario es un instrumento utilizado para la recogida de información, diseñado para poder cuantificar y universalizar la información y estandarizar el procedimiento de la entrevista. Su finalidad es conseguir la comparabilidad de la información[1].

En general, cuando se habla de cuestionarios se hace referencia a escalas de evaluación; por ejemplo, el cuestionario de calidad de vida SF-36 es una escala de evaluación. Por tanto, las escalas de evaluación son aquellos instrumentos o cuestionarios que permiten un escalamiento acumulativo de sus ítems, y que dan puntuaciones globales al final de la evaluación. Su carácter acumulativo las diferencia de los cuestionarios de recogida de datos, los inventarios de síntomas, las entrevistas estandarizadas o los formularios.

Tanto las entrevistas como los cuestionarios basan su información en la validez de la información verbal de percepciones, sentimientos, actitudes o conductas que transmite el encuestado; información que puede ser difícil de contrastar y de traducir a un sistema de medición, es decir, a una puntuación; esta dificultad es la causante de la complejidad de establecer la calidad de este tipo de instrumentos.

La utilización de las escalas de evaluación se basa en la psicofísica y la psicometría. La psicofísica aproxima el proceso de cuantificación de la percepción (trasladar a un sistema numérico

* Autor para correspondencia.
Correo electrónico: mjesus.garciadeyebenes@ser.es (M.J. García de Yébenes Prous).

1699-258X/$ - see front matter © 2008 Elsevier España, S.L. Todos los derechos reservados.
doi:10.1016/j.reuma.2008.09.007

fenómenos intangibles, como los síntomas o la discapacidad, mediante analogías). La psicometría permite estudiar la adecuación de la escala al fenómeno objeto de la medición y la calidad de la medida[1].

El desarrollo de un cuestionario es un proceso laborioso que puede llevar meses antes de conseguir una versión definitiva que satisfaga las expectativas previstas. Por esto, se debe tender a utilizar cuestionarios que ya hayan demostrado su utilidad en otros estudios, lo que, además, permite comparar resultados. Sin embargo, hay ocasiones en las que es inevitable diseñar nuevos instrumentos, por ejemplo, cuando los instrumentos existentes han mostrado resultados poco satisfactorios o han demostrado su eficacia en medios de aplicación distintos, o bien cuando no haya ningún cuestionario adecuado para medir lo que se pretende medir. En estas circunstancias se justifica el diseño de un nuevo cuestionario y la evaluación de su utilidad antes de su aplicación. Los cuestionarios son instrumentos diseñados para medir una serie de parámetros que, en muchas ocasiones, son conceptos teóricos o abstractos. Estos objetos de medición no directamente observables se denominan «constructos»[2].

Un cuestionario válido, como todo instrumento de medición, debe reunir las siguientes características[1]:

1. Ser sencillo, viable y aceptado por pacientes, usuarios e investigadores (viabilidad).
2. Ser fiable y preciso, es decir, con mediciones libres de error (fiabilidad).
3. Ser adecuado para el problema que se pretende medir (validez de contenido).
4. Reflejar la teoría subyacente en el fenómeno o concepto que se quiere medir (validez de constructo).
5. Ser capaz de medir cambios, tanto en los diferentes individuos como en la respuesta de un mismo individuo a través del tiempo (sensibilidad al cambio).

Mientras la fiabilidad y la validez son exigencias necesarias a todos los instrumentos, la importancia de otras características psicométricas depende del contexto. Por ejemplo, la sensibilidad al cambio es importante si el instrumento se aplica como medida de respuesta en los ensayos clínicos, pero no si se utiliza para un estudio sobre opiniones o actitudes acerca de una enfermedad[2].

El análisis de las características métricas del instrumento es un proceso complejo que implica la evaluación de la viabilidad, fiabilidad, validez y sensibilidad al cambio (tabla 1).

El objetivo de este trabajo es describir la metodología de estudio de la viabilidad, fiabilidad y validez de los cuestionarios como escalas o instrumentos de medición que permiten obtener y cuantificar datos con el fin de poder comparar información. El análisis de la sensibilidad al cambio no forma parte de este artículo.

Viabilidad

Los mejores instrumentos son inservibles si su aplicación resulta difícil, compleja o costosa. Características como el tiempo empleado en la cumplimentación, la sencillez y la amenidad del formato, el interés, la brevedad y la claridad de las preguntas, así como la facilidad de la corrección, el registro, la codificación y la interpretación de los resultados son aspectos relacionados con la viabilidad *(feasibility)*. Esta característica se estudia mediante la realización de un estudio piloto en un grupo de alrededor de 30 individuos y sus resultados pueden utilizarse para efectuar las modificaciones oportunas al instrumento de medición.

Fiabilidad

La fiabilidad *(reliability)* es el grado con el que un instrumento mide con precisión, sin error. La fiabilidad mide la proporción de variación en las mediciones que es debida a la diversidad de valores que adopta la variable y no es producto del error; es decir, la fiabilidad mide la proporción de la variancia total atribuible a diferencias verdaderas entre los sujetos[2,3]. Un instrumento fiable es preciso, es decir, proporciona mediciones libres de error. La variación que se debe a un error puede obedecer a 2 tipos de errores:

1. Sistemático o sesgo: error que se produce de forma sistemática. Por ejemplo, un evaluador puede puntuar siempre por debajo de los otros evaluadores.
2. Aleatorio: error que se produce por factores debidos al azar. Por ejemplo, por diferentes circunstancias, un evaluador puede dar algunas veces puntuaciones superiores y otras veces, puntuaciones inferiores a las correctas. El error aleatorio es el que más afecta la fiabilidad de un instrumento.

La fiabilidad de un instrumento se valora mediante la consistencia interna, la fiabilidad test-retest o intraobservador y la fiabilidad interobservador.

I. Consistencia interna

Esta propiedad se refiere a la coherencia de los componentes del instrumento de medición, es decir, se refiere a que los ítems que miden un mismo atributo presenten homogeneidad entre

Tabla 1
Características de los instrumentos de medición

Término	Sinónimo	Aspectos que se deben considerar	Técnica de análisis
Viabilidad	*Feasibility*	Tiempo empleado	Estudio piloto
		Claridad de preguntas	
		Registro, codificación	
		Interpretación de resultados	
Fiabilidad	*Reliability*	Consistencia interna	Alfa de Cronbach
		Intraobservador	CCI, índice kappa, método gráfico de Bland y Altman
		Interobservador	CCI, índice kappa, método gráfico de Bland y Altman
Validez	*Validity*	Lógica *(face validity)*	Redacción de las preguntas
		De contenido	Opinión de expertos
		De constructo	Constructo análisis factorial
		De criterio	Pruebas diagnósticas
Sensibilidad al cambio	*Responsiveness*	Intrínseca	En función del diseño y del tipo de cambio previsible
		Extrínseca	

CCI: coeficiente de correlación intraclase.

ellos. Una escala consistente garantiza que todos sus componentes o ítems midan un solo constructo que es homogéneo. Si la escala tiene una elevada consistencia interna, la suma de las puntuaciones puede representar la medición de un único constructo con el que, en general, mantiene una relación lineal.

Los cuestionarios se desarrollan para medir separadamente diferentes componentes o dimensiones de un problema. Por ejemplo, un cuestionario sobre salud puede estar dividido en preguntas sobre salud física y mental; se espera que haya una buena concordancia entre las distintas preguntas que miden un mismo componente. Por consiguiente, si un cuestionario está compuesto por diferentes subescalas, cada una de las cuales pretende medir una dimensión diferente del mismo fenómeno, debe evaluarse la consistencia interna de cada una de ellas[2,3]. La consistencia interna de una escala de valoración depende del número de ítems que componen el instrumento y de la correlación media entre ellos, y se evalúa en una única aplicación del instrumento mediante el método estadístico alfa de Cronbach[4], con valores comprendidos entre 0 y 1, y que se interpreta de forma similar a un coeficiente de correlación.

Por ejemplo, el índice AUSCAN (Australian/Canadian Osteoarthritis Hand Index 'Índice australiano/canadiense de la osteoartritis de mano') contiene 3 subescalas que evalúan el dolor (5 ítems), la rigidez (1 ítem) y la capacidad funcional (9 ítems) de pacientes con osteoartritis de las manos durante las 48 h previas. Las subescalas pueden utilizarse de forma individual o sumarse para obtener una única puntuación global. En un estudio se evaluó la consistencia interna de la puntuación global y de las subescalas. El alfa de Cronbach de la escala global fue de 0,96 mientras que los de las subescalas de dolor y de capacidad funcional fueron de 0,93 y 0,94, respectivamente[5].

II. Fiabilidad test-retest o intraobservador

La repetibilidad o fiabilidad test-retest se refiere a sí, cuando se administra un cuestionario a la misma población en 2 ocasiones diferentes en el tiempo, se obtienen resultados idénticos o similares; por tanto, mide la estabilidad de las puntuaciones otorgadas por el mismo evaluador en los mismos sujetos y con el mismo método en momentos diferentes. Esta técnica comporta dificultades prácticas. Por ejempo, si el tiempo transcurrido entre ambas aplicaciones es muy largo, el fenómeno medido puede presentar variaciones, mientras que si es demasiado breve puede haber un recuerdo de las respuestas (efecto de aprendizaje). En ambos casos se obtiene una medición distorsionada de la repetibilidad; además, algunos individuos pueden no aceptar una segunda aplicación del cuestionario. Sin embargo, este método es útil en variables bioquímicas o de laboratorio. Su análisis se realiza mediante el coeficiente de correlación intraclase (CCI) para escalas de medición cuantitativa y mediante el índice kappa de Cohen para escalas de medición cualitativa[6,7].

Una limitación importante del CCI es su dependencia de la variabilidad de los valores observados. Si los sujetos estudiados varían poco en sus puntuaciones (muestra homogénea), el CCI tiende a ser bajo, mientras que en muestras muy heterogéneas tiende a ser más elevado. Bland y Altman propusieron un método gráfico alternativo para evaluar la concordancia, de forma que el resultado no dependiera de la naturaleza del grupo de estudio. Sin embargo, la estimación del grado de acuerdo es subjetiva y no proporciona un índice objetivo como el CCI[8].

III. Fiabilidad interobservador

Se refiere al grado de acuerdo que hay entre 2 o más evaluadores que valoran a los mismos sujetos con el mismo instrumento. Los problemas más importantes en el análisis de esta dimensión de la fiabilidad son el error sistemático y la proporción de acuerdos que se debe al azar. Los métodos estadísticos más utilizados para su evaluación son los comentados en el apartado anterior.

En los últimos años la exploración ecográfica ha despertado un gran interés como método de evaluación de la actividad o de la respuesta terapéutica de diversas enfermedades reumáticas. En este sentido, Szkudlarek et al publicaron un estudio de fiabilidad interobservador de los hallazgos ultrasonográficos en articulaciones de los dedos de pacientes con artritis reumatoide que fueron evaluados por 2 investigadores con distinta formación. Se analizaron distintos parámetros (erosión ósea, engrosamiento de la membrana sinovial, derrame articular y señal Power Doppler) que se puntuaron en una escala semicuantitativa de 0 a 4, y también como presencia o ausencia de cada alteración. Se calculó la fiabilidad interobservador para cada parámetro mediante los 3 métodos estadísticos propuestos: CCI, índice kappa y método gráfico de Bland y Altman. El CCI y el índice kappa de los parámetros examinados mostraron una fiabilidad moderada o buena (0,61 a 0,81 y 0,48 a 0,68) con un acuerdo global elevado (del 79 al 91%)[9].

Validez

La validez de un instrumento se refiere a su capacidad para medir aquello para lo que ha sido diseñado. Al igual que en el caso de la fiabilidad, hay diferentes dimensiones de la validez de un instrumento: una dimensión lógica o aparente, una de contenido, una de constructo o concepto y una de criterio.

I. Validez lógica o aparente

La validez lógica o aparente se refiere al grado en que «parece» que un cuestionario mide lo que quiere medir a juicio de los expertos y de los propios sujetos. La decisión sobre si las preguntas deben tener o no validez lógica ha de tomarse antes de iniciar su redacción. Si las preguntas carecen de validez lógica es muy probable que los sujetos estudiados rechacen contestar las preguntas. No obstante, en algunos casos puede tener interés formular preguntas carentes de validez lógica. Por ejemplo, cuando se intenta abordar temas muy sensibles o conflictivos, la utilización de preguntas directas (con mucha validez lógica) puede hacer que el sujeto no conteste o falsee la respuesta, por lo que puede ser preferible realizar preguntas que aborden el tema de forma más indirecta, con menor validez lógica[2].

II. Validez de contenido

La validez de contenido es el grado en que la medición abarca la mayor cantidad de dimensiones del concepto que se quiere estudiar; por tanto, se considera que un instrumento es válido por su contenido si contempla todos los aspectos relacionados con el concepto en estudio. Esta dimensión de la validez se relaciona con la composición del instrumento y valora si éste contiene una muestra representativa (ítem) de los componentes del constructo que pretende medir. Supone el examen sistemático del contenido de la herramienta de medición para determinar si sus ítems son relevantes (si todos están relacionados con el concepto que se quiere medir) y representativos del dominio que se pretende medir (si representan las características esenciales del constructo y si están en las proporciones adecuadas).

La evaluación de la validez de contenido se basa en juicios de diferente procedencia (revisión de la literatura médica, opinión de expertos, estudios piloto). Este proceder debe garantizar, de forma empírica, que el contenido del instrumento sea adecuado.

Hay también otras formas de evaluar la validez de contenido, como el análisis factorial que explora las respuestas a las preguntas del cuestionario e intenta agruparlas en función de factores subyacentes que identifican las posibles dimensiones.

La diferencia entre la validez aparente y la validez de contenido reside en que la evaluación de esta última es un proceso más exhaustivo, y quizás más formal, en el que deberían participar tanto investigadores y médicos clínicos como miembros de la población diana.

III. Validez de constructo

Evalúa el grado en que el instrumento refleja la teoría del fenómeno o del concepto que se quiere medir. La validez de constructo garantiza que las mediciones que resulten de las respuestas del cuestionario puedan ser consideradas y utilizadas como medición del fenómeno estudiado. Se define, por tanto, como la capacidad de un instrumento para medir adecuadamente un constructo teórico. La medición de conceptos teóricos requiere una identificación previa del contenido de los instrumentos que se utilizarán y la elaboración de un modelo conceptual que ayude a interpretar los resultados obtenidos con estos instrumentos.

La validación de constructo representa el grado en que una medición se relaciona con otras mediciones de manera consistente con las hipótesis teóricas que definen el fenómeno o constructo que se quiere medir, y es una de las alternativas más frecuentes en caso de ausencia de un criterio de referencia o criterio externo[10].

Un método muy utilizado para evaluar la validez de constructo es el análisis factorial, que agrupa las respuestas en función de factores subyacentes; por lo que en estos casos se la denomina validez factorial. Mediante esta técnica, se analizan las interrelaciones existentes entre un conjunto de variables para intentar explicarlas a través de la extracción de los denominados factores.

Otro procedimiento más sencillo es examinar si el concepto en cuestión se relaciona con otras mediciones de forma consistente a lo esperable mediante análisis de regresión lineal o coeficientes de correlación (validez convergente)[2,10]. Por ejemplo, la valoración ecográfica de la inflamación sinovial ha demostrado validez de constructo, ya que en estudios transversales ha presentado buena concordancia con los índices clínicos de actividad inflamatoria y en estudios longitudinales se ha observado correlación entre los cambios sinoviales ecográficos tras tratamiento y los cambios clínicos y analíticos[11].

IV. Validez de criterio

En general, cuando se diseña un nuevo instrumento de medición se dispone de algún método alternativo de medición del fenómeno estudiado con validez demostrada, que se lo toma como referencia para determinar la validez del nuevo instrumento. Siempre que se disponga de un método de referencia adecuado se debe evaluar la validez de criterio del nuevo cuestionario. Cuando se habla de validar un cuestionario, los investigadores suelen referirse a la validez de criterio. El criterio externo o criterio de referencia debe ser una medición independiente, es decir, debe obtenerse por un método diferente en el que no intervengan los resultados del cuestionario.

Éste es el tipo de validez al que generalmente se hace referencia cuando se habla de validar un instrumento y debe seguir los siguientes pasos: a) identificar un criterio externo relevante y fiable; b) conseguir una muestra de sujetos representativa de la población en la que será usado el instrumento; c) administrar el instrumento y obtener una puntuación para cada sujeto, y d) evaluar a cada uno de los individuos con el criterio externo de referencia. El prototipo de la validez de criterio es el análisis de pruebas diagnósticas.

Análisis de pruebas diagnósticas

Se diseña un cuestionario o una escala para detectar la presencia o ausencia de un determinado proceso. La escala en cuestión se considera válida si clasifica a los sujetos según presenten o no el proceso con pocos errores. Por esta razón, es importante determinar el grado de similitud entre los resultados obtenidos en el cuestionario y los obtenidos de un criterio externo de referencia fiable y ampliamente aceptado como medida válida (siempre positivo en presencia del proceso y siempre negativo en ausencia del mismo) del diagnóstico de este proceso.

El criterio externo es un criterio dicotómico (presencia o ausencia de enfermedad), mientras que la escala del cuestionario es una medición continua. En estos casos hay que elegir un valor o un punto de corte a partir del que se considerará que la cifra obtenida constituye un resultado positivo en la escala. Al establecer este punto de corte se puede clasificar a los sujetos en sanos o enfermos según si el valor obtenido en la prueba es inferior o superior al del punto de corte o umbral elegido. La clasificación generada al elegir un determinado punto de corte comporta 2 tipos de errores: falsos positivos o sujetos sanos diagnosticados como enfermos y falsos negativos o sujetos enfermos diagnosticados como sanos[11,12].

El planteamiento del análisis de validez de una prueba diagnóstica se inicia a partir de la construcción de una tabla de 2×2 (tabla 2).

La validez de una prueba diagnóstica se evalúa mediante los índices de sensibilidad y de especificidad.

Sensibilidad

Se denomina sensibilidad (S) de una prueba diagnóstica a la proporción de individuos con la enfermedad que tienen un test positivo. Los test muy sensibles son aquellos que detectan a la mayoría de los individuos enfermos (pocos falsos negativos).

$$S = \frac{\text{Verdaderos positivos}}{\text{Total enfermos}} = \frac{VP}{VP + FN}$$

Especificidad

Se denomina especificidad (E) de una prueba diagnóstica a la proporción de individuos sin la enfermedad que tienen un resultado negativo en la prueba. Las pruebas más específicas son aquellas que descartan la enfermedad en la mayoría de los sujetos sanos (pocos falsos positivos).

$$E = \frac{\text{Verdaderos negativos}}{\text{Total no enfermos}} = \frac{VN}{VN + FP}$$

En general, se considera que la prueba diagnóstica tiene una validez aceptable si su sensibilidad y su especificidad son iguales o superiores a 0,80[2,12,13].

Tabla 2
Análisis básico de una prueba diagnóstica

Resultado de la prueba	Criterio externo de referencia No enfermo	Enfermo
Positivo	FP	VP
Negativo	VN	FN
TOTAL	FP+VN	VP+FN

FN: falso negativo; FP: falso positivo; VN: verdadero negativo; VP: verdadero positivo.

Cuando la prueba diagnóstica proporciona un resultado cuantitativo, la sensibilidad y la especificidad dependen del punto de corte elegido, es decir, del valor de la prueba a partir del que se considera que el sujeto presenta un resultado positivo o negativo en el test. La decisión del punto de corte debe ser cuidadosamente sopesada pues hay una interdependencia entre la sensibilidad y la especificidad, de forma que el incremento de una de ellas conlleva una disminución de la otra. A la hora de elegir el punto de corte se debe tener en cuenta el objetivo fundamental de la prueba.

Curvas de eficacia diagnóstica

Cuando los valores de una prueba diagnóstica siguen una escala cuantitativa, la sensibilidad y la especificidad varían según el punto de corte elegido para clasificar a la población como enferma o no enferma; es decir, son índices de la validez de la prueba diagnóstica para un determinado punto de corte. En esta situación, una medición global de la validez de la prueba para el conjunto de todos los posibles puntos de corte se obtiene mediante el uso de curvas ROC (*receiver operating characteristics* 'curva de eficacia diagnóstica') (fig. 1)[14]. Para construir la curva ROC es necesario calcular la sensibilidad y la especificidad para todos los posibles puntos de corte. La sensibilidad (S) o proporción de verdaderos positivos se sitúa en el eje de ordenadas (Y) y en el eje de abscisas se coloca el complementario de la especificidad (1-especificidad) o proporción de falsos positivos; la curva ROC se dibuja uniendo los pares de valores (1-E; S) correspondientes a cada punto de corte. El área bajo la curva (ABC) se define como la probabilidad de clasificar correctamente a un par de individuos (uno sano y uno enfermo) seleccionados al azar al aplicarles la prueba. Este tipo de gráfico permite valorar 2 situaciones extremas:

Una prueba con discriminación perfecta (S = 1; E = 1) estará representada por una curva ROC situada a los lados izquierdo y superior del gráfico.

Una prueba sin discriminación diagnóstica (la probabilidad de diagnosticar correctamente tanto a un sujeto sano como a uno enfermo será de 0,5; S = 0,5; E = 0,5) estará representada por la diagonal principal del gráfico.

La curva ROC facilita la elección del punto de corte. En general, si el coste de cometer un falso positivo es similar al de cometer un falso negativo, el mejor punto de corte es el más próximo al ángulo superior izquierdo del gráfico[14].

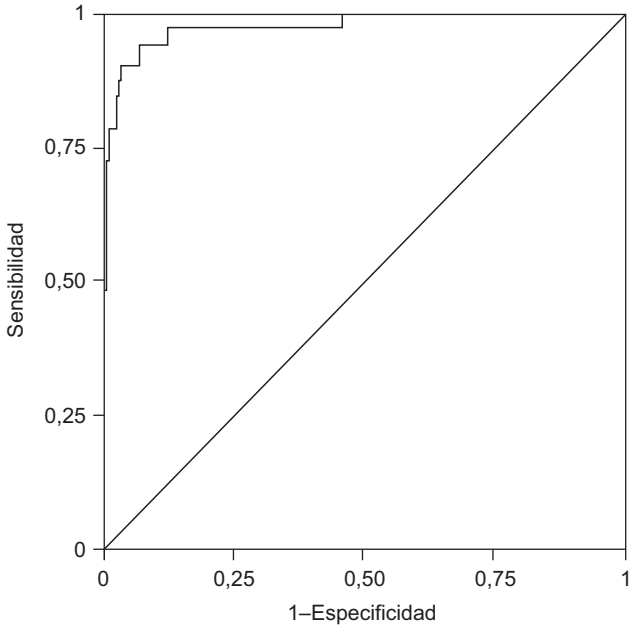

Figura 1. Curva de eficacia diagnóstica.

Comportamiento de una prueba diagnóstica

Además del estudio de la validez de una prueba diagnóstica, es importante evaluar su comportamiento cuando se aplica en diferentes contextos clínicos. Para esto, es preciso calcular los valores predictivos y la eficiencia de la prueba:

Valor predictivo positivo

Es la proporción de sujetos con la enfermedad en el conjunto de individuos con resultado positivo en la prueba. Es decir, es la probabilidad de que un individuo con resultado positivo tenga la enfermedad.

$$\text{VPP} = \frac{\text{Verdaderos positivos}}{\text{Total positivos}} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Valor predictivo negativo

Es la proporción de sujetos sin la enfermedad en el conjunto de individuos con resultado negativo en la prueba. Es decir, es la probabilidad de que un individuo con resultado negativo no tenga la enfermedad.

$$\text{VPN} = \frac{\text{Verdaderos negativos}}{\text{Total negativos}} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Valor global o eficiencia

Es la proporción total de sujetos clasificados correctamente.

$$\text{VG} = \frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{Total sujetos}} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}$$

Hay que tener en cuenta que los valores predictivos, tanto positivos como negativos, son índices que dependen de la prevalencia o de la probabilidad previa de la enfermedad, es decir, evalúan el comportamiento de la prueba diagnóstica en una población con una determinada proporción de sujetos sanos. La prevalencia es el factor más determinante de los valores predictivos. Al ser características intrínsecas de una medición, la sensibilidad y la especificidad no experimentan grandes variaciones según el lugar en el que se apliquen, siempre que se realice en condiciones similares. Por esta razón, la predictividad de una medición no se puede evaluar sin considerar la prevalencia de la enfermedad; si es alta, un resultado positivo tiende a confirmar su presencia, mientras que si el resultado es negativo no ayudará a excluirla. Por el contrario, cuando la prevalencia es baja, un resultado negativo permite descartar la enfermedad con un elevado margen de confianza, pero no permite afirmar su existencia. En general, el valor predictivo positivo disminuye a medida que la prueba diagnóstica se aplica a poblaciones con prevalencia de enfermedad más baja. Esto se debe a que una prueba que produce falsos positivos se aplica a una población de sujetos mayoritariamente sanos, por lo que en esta situación es relativamente fácil obtener muchos falsos positivos y, por tanto, el valor predictivo para positivos disminuye[15].

Razones de probabilidad

Una forma de evitar la influencia de la prevalencia en la validez de una prueba diagnóstica es la utilización de las llamadas razones de verosimilitud *(likelihood ratios)* que relacionan la

Tabla 3
Preguntas para evaluar la validez de un cuestionario

Concepto	Pregunta	Previsible en una escala de fatiga adecuada
Validez		
Validez aparente	¿El método parece sensible?	El lenguaje refleja las ideas del paciente sobre la fatiga
Validez de contenido	¿El origen de las preguntas se basa en los pacientes?	Los pacientes son el origen o los revisores de las preguntas
	¿Se han incluido todos los ítems necesarios?	P. ej., aspectos físicos, emocionales, cognitivos, gravedad, consecuencias
	¿Se han evitado todos los ítems confusos?	Ítems que se podrían confundir con discapacidad
Validez de criterio	¿Se ha comparado con un criterio externo de fatiga?	Evaluado frente a otra escala de fatiga
Validez de constructo	¿Converge con otras variables adecuadas?	P. ej., correlación moderada con dolor, inflamación, estado de ánimo, anemia
Fiabilidad		
Consistencia interna	¿Es internamente consistente?	Correlación interítem moderada o elevada
Estabilidad	¿La escala es estable?	No se modifica en pacientes estables
Viabilidad	¿Cuánto tiempo tarda en cumplimentarse?	10 a 15 min máximo
	¿Es autorreferida o administrada por un entrevistador?	Es adecuado que las escalas subjetivas sean autorreferidas
	¿Es fácil de puntuar e interpretar?	Instrucciones claras

sensibilidad y la especificidad en un solo índice, por lo que no varían con la prevalencia del proceso.

Razón de verosimilitud para un resultado positivo

Se calcula al dividir la proporción de sujetos enfermos con resultado positivo en la prueba (sensibilidad) por la proporción de sujetos no enfermos, pero cuyo resultado también ha sido positivo (1-especificidad).

$$\mathrm{RV}+ = \frac{\text{Sensibilidad}}{1 - \text{Especificidad}}$$

Razón de verosimilitud para un resultado negativo

Es el cociente entre el complementario de la sensibilidad y la especificidad.

$$\mathrm{RV}- = \frac{1 - \text{Sensibilidad}}{\text{Especificidad}}$$

De estos 2 índices, el más utilizado en la práctica es la razón de verosimilitud para un resultado positivo, por lo que se lo conoce con la denominación genérica de «razón de verosimilitud». Si, por ejemplo, se obtiene una razón de verosimilitud de 8, este valor indica que en el grupo de enfermos la probabilidad de encontrar un resultado positivo en la prueba es 8 veces superior que en el grupo de no enfermos.

Se considera que la razón de verosimilitud es una medida más para valorar la validez de una prueba diagnóstica. Este índice cuenta con la ventaja de relacionar la sensibilidad y la especificidad en una sola medición y, por tanto, es independiente de la prevalencia del proceso. Otra utilidad de la razón de verosimilitud es que también permite el cálculo de los valores predictivos.

Con el fin de calcular la sensibilidad y la especificidad del eco Doppler en el diagnóstico de la artritis de manos y muñecas y definir un punto de corte para 2 índices de inflamación (el índice de resistencia y la fracción de color), Terslev et al realizaron un estudio sobre una muestra de 88 pacientes con artritis reumatoide activa y 27 controles sanos. Todos los individuos de la muestra fueron estudiados con eco Doppler para calcular el índice de resistencia y la fracción de color de las articulaciones de la muñeca (metacarpofalángicas e interfalángicas proximales). Se construyeron curvas ROC para ambos parámetros de inflamación y se seleccionaron los puntos de corte con mayor sensibilidad y especificidad. El área bajo la curva fue de 0,84 para ambos índices. El punto de corte para la fracción de color fue de 0,01 con valores de sensibilidad y especificidad de 0,92 y 0,73, respectivamente. En el caso del índice de resistencia, se eligió como punto de corte el valor 0,83 con sensibilidad y especificidad de 0,72 y de 0,70, respectivamente. Los autores concluyeron que el eco Doppler puede detectar vascularización de la membrana sinovial inflamada con elevada sensibilidad y moderada especificidad[16].

Conclusión

La utilización de instrumentos de medición inadecuados o no válidos puede producir resultados no fiables o confusos. En una revisión sistemática de diferentes escalas de medición de la fatiga en la artritis reumatoide, los autores encontraron que únicamente 6 de 23 escalas presentaban pruebas de validez razonables. Los autores desarrollaron una serie de preguntas para evaluar la validez de un cuestionario que puede resultar de gran utilidad y que se presentan en la tabla 3[17].

Bibliografía

1. Martín Arribas MC. Diseño y validación de cuestionarios. Matronas profesión [serial online] 2004 [consultado 19/5/2008]; 5:23–9. Disponible en: http://www.enferpro.com/documentos/validacion_cuestionarioswww.enferpro.com/documentos/validacion_cuestionarios.
2. Argimón Pallás JM, Jiménez Vila J. Métodos de investigación clínica y epidemiológica. 3ª ed. Madrid: Ediciones Harcourt; 2006.
3. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis. 1985;38:27–36.
4. Altman DG, Bland JM. Cronbach's alpha. BMJ. 1997;314:572.
5. Allen KD, Jordan JM, Renner JB, Kraus VB. Validity, factor structure and clinical relevance of the AUSCAN Osteoarthritis hand index. Arthritis Rheum. 2006;54:551–6.
6. Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase. Med Clin. 1998;110:142–5.
7. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med. 1994;13:2465–76.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.
9. Szkudlarek M, Court-Payen M, Jacobsen S, Klarlund M, Thomsen HS, Ostergaard M. Interobserver agreement in ultrasonography of the finger and toe joints in rheumatoid arthritis. Arthritis Rheum. 2003;48:955–62.
10. Van der Hofstadt CJ, Rodríguez-Marin J. Adaptación de un cuestionario para la medida de la representación de la enfermedad. Psicothema [serial online] 1997 [citado 22 May 2008]; 9:237–45. Disponible en: URL: www.psicothema.com.

11. Naredo E. Evaluación de la artritis reumatoide por técnicas de imagen: ecografía. Reumatol Clin. 2006;2(Supl 2):S13–7.
12. Pozo Rodríguez F. La eficacia de las pruebas diagnósticas (I). Med Clin (Barc). 1988;90:779–85.
13. Pozo Rodríguez F. La eficacia de las pruebas diagnósticas (II). Med Clín (Barc). 1988;91:177–83.
14. Bargueño MJ, García-Bastos JL, González-Buitrago JM. Las curvas ROC en la evaluación de las pruebas diagnósticas. Med Clín (Barc). 1995;104:661–70.
15. Cabello López JB, Pozo Rodríguez F. Métodos de investigación en Cardiología Clínica (X). Estudios de evaluación de las pruebas diagnósticas en Cardiología. Rev Esp Cardiol. 1997;50:507–19.
16. Terslev L, Von der Recke P, Torp-Pedersen S, Koenig MJ, Bliddal H. Diagnostic sensitivity and specificity of Doppler ultrasound in rheumatoid arthritis. J Rheumatol. 2008;35:8–10.
17. Hewlett S, Mehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: A systematic review of scales in use. Arthritis Care Res. 2007;57:429–39.