EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Fundamentos de medicina basada en la evidencia

Evaluación de la precisión de las pruebas diagnósticas (1). Variables discretas

Ochoa Sangrador C[1], Molina Arias M[2]
[1]Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.
[2]Servicio de Gastroenterología. Hospital Infantil Universitario La Paz. Madrid. España.

Correspondencia: Carlos Ochoa Sangrador, cochoas2@gmail.com

Palabras clave en español: precisión, reproducibilidad, pruebas diagnósticas, fiabilidad de las pruebas diagnósticas, variación intra-interobservador, estadístico kappa.

Fecha de recepción: 20 de mayo de 2017 • **Fecha de aceptación:** 24 de mayo de 2017
Fecha de publicación del artículo: 31 de mayo de 2017

Evid Pediatr. 2017;13:28.

CÓMO CITAR ESTE ARTÍCULO

Ochoa Sangrador C, Molina Arias M. Evaluación de la precisión de las pruebas diagnósticas (1). Variables discretas. Evid Pediatr. 2017;13:28.

Este artículo está disponible en: http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2017;13:28.
 • ISSN: 1885-7388

Evaluación de la precisión de las pruebas diagnósticas (1). Variables discretas

Ochoa Sangrador C[1], Molina Arias M[2]

[1]*Servicio de Pediatría. Hospital Virgen de la Concha. Zamora. España.*

[2]*Servicio de Gastroenterología. Hospital Infantil Universitario La Paz. Madrid. España.*

Correspondencia: Carlos Ochoa Sangrador, cochoas2@gmail.com

INTRODUCCIÓN

En documentos previos de esta serie hemos abordado cómo evaluar la validez de una prueba diagnóstica respecto a un patrón de referencia. Si una prueba mide realmente lo que queremos medir, la consideramos lo suficientemente válida como para confiar en sus resultados, porque hemos comprobado que concuerdan con los de pruebas más agresivas, caras o no disponibles, o bien con la confirmación clínica del diagnóstico, tras comprobar la evolución del paciente[1].

Sin embargo, la confianza que asignamos a una prueba diagnóstica no depende solo de su validez, también depende de su precisión o fiabilidad, esto es, de la estabilidad que muestran sus mediciones cuando se repiten en condiciones similares. La fiabilidad es un requisito previo al de validez, ya que es necesario saber que una prueba es capaz de medir "algo", antes de plantearse contrastar su validez. Si mediciones repetidas de una característica con un mismo instrumento son inconsistentes, la información resultante no va a poder aportar nada al diagnóstico. No obstante, una prueba muy fiable en sus mediciones, pero en la que estas no sean válidas, tampoco tiene ninguna utilidad.

La fiabilidad o precisión de una prueba es su capacidad para producir los mismos resultados cada vez que se aplica en similares condiciones. La fiabilidad implica falta de variabilidad. Sin embargo, las mediciones realizadas por las pruebas diagnósticas están sujetas a múltiples fuentes de variabilidad. Esta variabilidad puede encontrarse en el propio sujeto objeto de la medición (variabilidad biológica), en el instrumento de medida propiamente dicho o en el observador que la ejecuta o interpreta. A la hora de analizar y controlar la fiabilidad de las pruebas diagnósticas tiene especial interés estudiar la variabilidad encontrada entre las mediciones realizadas por dos o más observadores o instrumentos y la variabilidad encontrada entre mediciones repetidas realizadas por el mismo observador o instrumento.

Existen diversos métodos para la valoración de la fiabilidad de las mediciones clínicas. Los más adecuados en función del tipo de dato a medir son los siguientes: 1) índice kappa, para datos discretos nominales; 2) índice kappa ponderado, para resultados discretos ordinales, y 3) desviación estándar intrasujetos, coeficiente de correlación intraclase y método de Bland-Altman para datos continuos. En este primer documento abordaremos los métodos para variables discretas.

VARIABLES DISCRETAS NOMINALES. ÍNDICE KAPPA

El índice kappa puede aplicarse a pruebas cuyos resultados solo tengan dos categorías posibles o más de dos sin un orden jerárquico entre ellas. En la tabla 1 se presentan los resultados de un estudio en el que dos médicos evaluaron, de forma ciega, las radiografías de tórax de 100 niños con sospecha de neumonía (datos figurados). La tabla de contingencia refleja los recuentos de casos en que hay acuerdo (casillas a y d) y desacuerdo (casillas b y c).

La forma más sencilla de expresar la concordancia entre las dos evaluaciones es mediante el porcentaje o proporción de acuerdo o concordancia simple (P_o), que corresponde a la proporción de observaciones concordantes:

$$P_o = \frac{a + d}{Total} = \frac{4 + 80}{100} = 0{,}84\ (84\%)$$

Una concordancia del 84% podría ser interpretada como buena; sin embargo, es preciso tener en cuenta que parte del acuerdo encontrado puede ser debido al azar (si el médico sabe que solo uno de cada diez pacientes con sospecha de neumonía la tiene, ajustará consciente o inconscientemente sus diagnósticos a esa frecuencia). Las observaciones esperadas por azar en cada casilla de la tabla de contingencia se pueden calcular a partir del producto de los marginales de la fila y columna correspondientes, dividido por el total. En la tabla 2 se presentan los cálculos para cada una de las casillas

Tabla 1. Evaluación por parte de dos médicos de las radiografías de tórax de 100 niño con sospecha de neumonía (datos figurados). Las casillas reflejan el recuento de casos en que hay acuerdo y desacuerdo

		Médico A		
		Neumonía	No	
Médico B	Neumonía	4	6	10
	No	10	80	90
		14	86	100

Tabla 2. Estimación de las observaciones esperadas por azar en la tabla de contingencia del ejemplo de la tabla 1

		Médico A		
		Neumonía	No	
Médico B	Neumonía	$a' = \frac{10 \times 14}{100} = 1,4$	$b' = \frac{10 \times 86}{100} = 8,6$	10
	No	$c' = \frac{90 \times 14}{100} = 12,6$	$d' = \frac{90 \times 86}{100} = 77,4$	90
		14	86	100

del ejemplo de la tabla 1. Considerando estos recuentos estimados, la proporción de acuerdo esperada por azar sería:

$$P_e = \frac{a' + d'}{N} = \frac{\frac{10 \times 14}{100} + \frac{90 \times 86}{100}}{100} = \frac{1,4 + 77,4}{100} = 0,79\ (79\%)$$

Podemos constatar que existe acuerdo por azar en una elevada proporción de observaciones (79%). Si excluimos del análisis dichas observaciones, solo quedarán cinco observaciones concordantes (84-79=5) en un total de 21 observaciones (100-79=21), lo que supone un grado de acuerdo no debido al azar del 24% (5/21=0,24). Si formulamos este cálculo como probabilidades en vez de recuentos obtendremos el índice kappa.

El índice kappa nos ofrece una estimación del grado de acuerdo no debido al azar a partir de la proporción de acuerdo observado (P_o) y la proporción de acuerdo esperado (P_e):

$$\kappa = \frac{P_o + P_e}{1 - P_e}$$

Aplicando esta fórmula en nuestro ejemplo (tabla 1) obtenemos:

$$\kappa = \frac{P_o + P_e}{1 - P_e} = \frac{0,84 - 0,75}{1 - 0,75} = 0,36,$$

lo que supone un grado de concordancia no debido al azar del 36%, considerablemente más bajo que la proporción de acuerdo observado.

El índice kappa puede adoptar valores entre -1 y 1. Es 1 si existe un acuerdo total, 0 si el acuerdo observado es igual al esperado y menor de 0 si el acuerdo observado es inferior al esperado por azar. La interpretación más aceptada de los rangos de valores situados entre 0 y 1 se expone en la tabla 3[2,3]. Al igual que otros estimadores poblacionales, los índices kappa se deben calcular con sus intervalos de confianza[3].

El índice kappa también puede ser aplicado a pruebas cuyos resultados tengan más de dos categorías nominales, utilizando la misma metodología para el cálculo del acuerdo esperado por azar.

Tabla 3. Interpretación de los valores del índice kappa

Valor de kappa	Grado de concordancia
0,81-1,00	Excelente
0,61-0,80	Buena
0,41-0,60	Moderada
0,21-0,40	Ligera
≤ 0,20	Mala

VARIABLES DISCRETAS ORDINALES. ÍNDICE KAPPA PONDERADO

El índice kappa ponderado debe emplearse cuando el resultado de la prueba analizada puede adoptar más de dos categorías, entre las que existe cierto orden jerárquico (resultados discretos ordinales). En esta situación, pueden existir distintos grados de acuerdo o desacuerdo entre las evaluaciones repetidas. Veamos un ejemplo. En la tabla 4 se presentan los resultados de dos evaluaciones sucesivas de un cuestionario (test-retest), diseñado para detectar el consumo problemático de alcohol en adolescentes (datos figurados). Los resultados se expresan en tres categorías: riesgo bajo, medio y alto. Es evidente que no puede considerarse igual una discrepancia entre riesgo bajo y medio que entre bajo y alto.

El índice kappa ponderado nos permite estimar el grado de acuerdo, considerando de forma diferente esas discrepancias. Para ello, debemos asignar diferentes pesos a cada nivel de concordancia. Habitualmente se asignará un peso 1 al acuerdo total (100% de acuerdo) y un peso 0 al desacuerdo extremo. A los desacuerdos intermedios se les asignarán pesos intermedios, en función del significado que tengan las distintas discordancias en el atributo estudiado. Así, si en nuestro ejemplo hemos optado por asignar un peso de 0,25 a las discordancias riesgo alto-medio, ello significa que cuando una de las evaluaciones clasifica el riesgo como alto y la otra como medio, el grado de acuerdo entre ambas es solo del 25%.

El índice kappa ponderado se calcula de forma similar al índice kappa, con la diferencia de que, en las fórmulas de las proporciones de acuerdo observado y esperado, las frecuencias de las

Tabla 4. Resultados de dos evaluaciones sucesivas, separadas por un corto periodo de tiempo (test-retest), de un cuestionario diseñado para detectar el consumo problemático de alcohol, en 100 adolescentes (datos figurados). Los resultados se expresan en tres categorías: riesgo bajo, medio y alto. Las casillas reflejan el recuento de casos en que hay acuerdo y desacuerdo

		1.ª evaluación			
		Riesgo bajo	Riesgo medio	Riesgo alto	
2.ª evaluación	Riesgo bajo	35	12	5	52
	Riesgo medio	8	10	5	23
	Riesgo alto	5	9	11	25
		48	31	21	100

distintas casillas se deben multiplicar por sus pesos respectivos. En la tabla 5 podemos ver los pesos asignados en el ejemplo de la tabla 4 y los cálculos de las observaciones esperadas por azar en cada casilla. Las proporciones de acuerdo observado (P_o), esperado (P_e) y el índice kappa ponderado (κ_w) para este ejemplo serán las siguientes (P_o y P_e calculados respectivamente con los valores de las tablas 4 y 5):

$$P_o = \frac{1 \times (35+10+11) + 0{,}25 \times (8+9+12+5)}{100} = 0{,}64$$

$$P_e = \frac{1 \times (24{,}9+7{,}1+5{,}2) + 0{,}25 \times (16{,}1+4{,}8+11+7{,}7)}{100} = 0{,}47,$$

$$\kappa_W = \frac{P_o + P_e}{1 - P_e} = \frac{0{,}64 - 0{,}47}{1-0{,}47} = 0{,}32$$

Es preciso señalar que las estimaciones de concordancia pueden variar de forma importante en función de los pesos elegidos. Una forma de estandarizar estos índices cuando no tenemos una hipótesis clara del grado de discordancia es utilizar un sistema de ponderación proporcional a la distancia entre categorías: los pesos bicuadrados. A cada casilla se le asigna un peso ($w_{i,j}$) igual a:

$$W_{ij} = 1 - \left(\frac{i-j}{\kappa - 1}\right)^2,$$

donde i es el número de columna en la tabla de contingencia, j el número de fila y k el número total de categorías (ver tabla 6). Los pesos bicuadrados, calculados con esta fórmula, de los acuerdos intermedios de nuestro ejemplo (alto-medio y medio-bajo) serían de 0,75.

Es interesante señalar que si se emplean estos pesos el valor del índice kappa ponderado se aproxima al del coeficiente de correlación intraclase, que veremos en un próximo documento de esta serie, cuando revisemos las medidas de concordancia para variables continuas.

Tabla 5. Pesos asignados a los distintos grados de acuerdo entre evaluaciones (en negrita en la esquina superior derecha de cada casilla) y recuentos esperados por azar en cada una de las casillas de la tabla 4 (ecuaciones de cada casilla)

		1.ª evaluación			
		Riesgo bajo	Riesgo medio	Riesgo alto	
2.ª evaluación	Riesgo bajo	**1** $\frac{52 \times 48}{100} = 24{,}9$	**0,25** $\frac{52 \times 31}{100} = 16{,}1$	**0** $\frac{52 \times 21}{100} = 10{,}9$	52
	Riesgo medio	**0,25** $\frac{23 \times 48}{100} = 11{,}0$	**1** $\frac{23 \times 31}{100} = 7{,}1$	**0,25** $\frac{23 \times 21}{100} = 4{,}8$	23
	Riesgo alto	**0** $\frac{25 \times 48}{100} = 12{,}0$	**0,25** $\frac{25 \times 31}{100} = 7{,}7$	**1** $\frac{25 \times 21}{100} = 5{,}2$	25
		48	31	21	100

Tabla 6. Pesos bicuadrados (en negrita) según el grado de concordancia

		1ª evaluación (κ=3 categorías)		
		Riesgo bajo i = 1	Riesgo medio i = 2	Riesgo alto i = 3
2.ª evaluación (κ =3 categorías)	Riesgo bajo j = 1	$1-\left(\frac{1-1}{3-1}\right)^2=1$	$1-\left(\frac{2-1}{3-1}\right)^2=0{,}75$	$1-\left(\frac{3-1}{3-1}\right)^2=0$
	Riesgo medio j = 2	$1-\left(\frac{1-2}{3-1}\right)^2=0{,}75$	$1-\left(\frac{2-2}{3-1}\right)^2=1$	$1-\left(\frac{3-2}{3-1}\right)^2=0{,}75$
	Riesgo alto j = 3	$1-\left(\frac{1-3}{3-1}\right)^2=0$	$1-\left(\frac{2-3}{3-1}\right)^2=0{,}75$	$1-\left(\frac{3-3}{3-1}\right)^2=1$

BIBLIOGRAFÍA

1. Ochoa Sangrador C, Orejas G. Epidemiología y metodología científica aplicada a la Pediatría (IV): pruebas diagnósticas. An Esp Pediatr. 1999;50:301-14.

2. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

3. Fleiss JL. The measurement of interrater agreement. En: Fleiss JL (ed.). Statistical methods for rates and proportions. Toronto: John Wiley & Sons; 1981. p. 212-36.