PUESTA AL DÍA

Métodos de investigación en cardiología clínica (IV)

Las mediciones clínicas en cardiología: validez y errores de medición

Jaime Latour[1], Víctor Abraira[2], Juan B. Cabello[3] y Javier López Sánchez[2]

[1]Institut Valencià d'Estudis en Salut Pública (IVESP). Valencia. [2]Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid. [3]Unidad de Investigación. Hospital General Universitari d'Alacant. Alicante.

bioestadística/ investigación biomédica/ escalas de valoración

La medición constituye una pieza clave de la actividad clínica. Frecuentemente, sin embargo, se comprueba la existencia de discrepancias entre las mediciones efectuadas por distintos clínicos o por el mismo clínico en 2 ocasiones diferentes. El origen de esa variabilidad se puede encontrar en el propio sujeto objeto de la medición (el paciente), en el aparato de medida propiamente dicho, o en el observador. La calidad de una medición se comprueba normalmente evaluando su reproducibilidad y su validez. La reproducibilidad se evalúa básicamente examinando la concordancia entre observadores, la concordancia intraobservadores y la concordancia test-retest. Los parámetros utilizados para medirla (coeficiente de correlación intraclase, coeficiente kappa, métodos gráficos, etc.) dependen del tipo de variable que se desea medir. La validez de la medición indica en qué medida la medición mide realmente lo que queremos medir. Cuando existe una prueba de referencia, la validez se estima mediante su comparación con el test (validez de criterio); cuando no existe una prueba de referencia aceptable se recurre a otras formas de validación que utilizan criterios subjetivos (validez de contenido y aparente) o empíricos (validez de constructo).

CLINICAL MEASUREMENT IN CARDIOLOGY: VALIDITY AND MEASUREMENT ERROR

Measurements represent an essential part of clinical activity. Very often, however, relevant disagreement in clinical measurements becomes apparent. The sources of this variability are the subjects (patients) that are measured, the measurement instrument itself, and the observer. The assessment of the quality of measurement usually relies on the evaluation of its reproducibility and its validity. The reproducibility is basically measured as the inter-observer concordance, the intra-observer concordance, and the test-retest concordance. The specific parameter used to its quantification (intra-class correlation coefficient, kappa index, graphic methods, etc.) depend on the kind of variable to be measured. The validity of the measurement is the degree to which the measurement is really measuring what we think it should. If an acceptable standard is available, then so called criterion validity is usually assessed. Otherwise the validity should be assessed by other ways that use subjective criteria (content validity and face validity) or empirical criteria (construct validity).

(*Rev Esp Cardiol* 1997; 50: 117-128)

INTRODUCCIÓN

La medición constituye una actividad omnipresente tanto en la práctica como en la investigación clínica. Como ejemplos de medición podemos citar desde actividades relativamente simples, como el registro de la presión arterial mediante un esfigmomanómetro, hasta actividades más sofisticadas como la determinación del nivel de creatincinasa en el suero, la cuantificación de la fracción de eyección por angiografía biplana o la evaluación de la calidad de vida de un enfermo coronario. En unos casos el proceso de medición se limita

Este artículo ha sido financiado, en parte, con las ayudas BISC 96/4782 y FIS 96/0421.

Correspondencia: Dr. V. Abraira.
Unidad de Bioestadística Clínica. Hospital Ramón y Cajal.
Ctra. Colmenar, km 9. 28034 Madrid.

a diferenciar grupos (p. ej., enfermos/sanos); otras veces se trata de un proceso de recuento (p. ej., el número de vasos coronarios con lesiones significativas); finalmente, algunas mediciones utilizan variables ordinales como el grupo de la NYHA, continuas (como la presión arterial) o una razón (como la fracción de acortamiento).

La medición *en el entorno del laboratorio* se ocupa habitualmente de fenómenos objetivos. En este contexto, la evaluación del error de medición se realiza a partir de unas pautas perfectamente estandarizadas, que incluyen la comprobación de 2 propiedades de la medición: su reproducibilidad (*reproducibility*) –o sea, su capacidad de dar el mismo resultado cuando se repite la medición sobre el mismo objeto– y su exactitud (*accuracy*) –es decir, la coincidencia de la medición con un patrón de referencia[1].

El problema de la calidad de la *medición en clínica* es más complejo. Por un lado, las condiciones en las que se realiza la medición no suelen ser perfectamente controlables (variabilidad fisiológica, falta de colaboración del paciente, etc.). En segundo lugar, el proceso de medición se puede ver muy influenciado por la subjetividad del observador. Finalmente, con frecuencia no existe un patrón de referencia aceptable para comprobar la exactitud de la medición. Estas dificultades añadidas explican la enorme variabilidad que presentan las mediciones clínicas[2]. El problema es especialmente grave a la hora de medir variables «frágiles» como el dolor o constructos complejos como la calidad de vida. Sin embargo, las discrepancias afectan también a la medición de variables clínicas habitualmente consideradas como «objetivas», tales como la medición del segmento de ST o la determinación de la fracción de eyección por ecocardiografía.

La calidad de las mediciones condiciona no sólo la calidad de la investigación, sino también la calidad de las decisiones clínicas que se apoyan en dichas mediciones[3]. Por todo ello, es aconsejable que el clínico conozca algunos fundamentos de la teoría de la medición, así como las causas más frecuentes de los errores de medida.

CAUSAS DE LA VARIABILIDAD EN LAS MEDICIONES CLÍNICAS

Cuando se realiza un estudio sobre concordancia en mediciones clínicas el objetivo principal no suele ser la comprobación de que existe variabilidad, sino la identificación de las causas de las discrepancias, para intentar corregirlas. Para ello puede ser útil separar tres fuentes de variabilidad. En primer lugar, la variabilidad real, debida a los sujetos en estudio, que es en definitiva lo que tratamos de medir. Para ello, seleccionamos un instrumento que nos ayude a distinguir entre los sujetos, diferenciando, por ejemplo, a un paciente normotenso de uno hipertenso. En segundo lugar, hay que considerar la variabilidad debida al propio procedimiento de medición. Por último, una parte de la variabilidad es atribuible al observador o usuario de la medición[4,5].

La importancia relativa de estas tres fuentes de variabilidad depende del tipo de variable que se está midiendo. Cuando se trata de variables simples, la variabilidad procede fundamentalmente de los datos de los sujetos o rasgo en estudio y el observador desempeña un papel secundario. Cuando la variable es más compleja, en cambio, la variabilidad depende en mayor medida del proceso de elaboración de esos datos, y el método y el observador pasan a tener un papel relevante.

Las discrepancias debidas al *paciente* se pueden deber a la variación biológica del sistema examinado. Por ejemplo, cualquier clínico sabe que la presión arterial o la frecuencia del pulso varían de hora a hora, en relación con múltiples factores (dieta, estrés, postura, etc.). En ocasiones, sin embargo, se ignora que esta variabilidad es aplicable, asimismo, a mediciones habitualmente consideradas como «exactas» tales como el ECG o la presión telediastólica del ventrículo izquierdo. El efecto de la medicación o la propia enfermedad pueden producir variaciones en los parámetros que deseamos medir, induciendo a discordancias entre observadores. Finalmente, la reelaboración de los datos de la historia clínica por parte del paciente, como ocurre cuando se insiste una y otra vez en la anamnesis, puede condicionar importantes inconsistencias. Este fenómeno lo experimenta con frecuencia (y de forma dolorosa) el médico en período de formación cuando, en presencia del jefe de servicio, el paciente relata una historia clínica radicalmente diferente de la que previamente le había referido al realizar la anamnesis.

Las inconsistencias debidas al *método* dependen a veces del ambiente en que se realiza la medición. La importancia de un entorno silencioso para detectar un suave soplo diastólico, o la necesidad de mantener un ambiente de privacidad para recoger datos sensibles de la anamnesis no requieren mayores comentarios. La falta de una buena interacción entre el clínico y el paciente dificulta con frecuencia la identificación de determinados datos de la historia clínica y compromete su consistencia. Por último, una utilización inapropiada de la herramienta diagnóstica, por ejemplo, una mala calibración del esfigmomanómetro, o una selección inadecuada del manguito pueden condicionar errores en la medición de la presión arterial.

El efecto del *observador* queda patente en la variabilidad biológica de los sentidos; por ejemplo, la concordancia en la auscultación cardíaca entre dos clínicos empeora probablemente cuando uno de ellos acaba de pasar una noche de guardia en vela. Otras veces las discrepancias no radican en la evidencia sen-

sorial, sino en la interpretación que se realiza de dicha evidencia. Este tipo de discrepancias se ve facilitado por la tendencia a registrar inferencias en lugar de evidencias; por ejemplo, cuando se describe un dolor como «pleurítico» en lugar de como un «dolor torácico inspiratorio». En ocasiones, las discrepancias se deben a la aplicación de diferentes criterios diagnósticos: por ejemplo, la identificación de un paciente como hipercolesterolémico dependerá del valor de colesterol que se considere como normal, y que puede ser diferente entre los distintos observadores. Una causa adicional de discrepancias debidas al observador se debe a las expectativas diagnósticas; por ejemplo, si un radiólogo conoce que el paciente tiene disnea, crepitantes basales y galope ventricular estará probablemente más predispuesto a establecer un diagnóstico de insuficiencia cardíaca que un radiólogo que carezca de datos clínicos.

VALIDEZ Y FIABILIDAD

La calidad de un instrumento de medida depende básicamente de dos propiedades: su fiabilidad (*reliability*) y su validez (*validity*).

El concepto de fiabilidad es elusivo. Cuando se emplea en ámbitos clínicos, este término hace referencia habitualmente a la estabilidad de la medida cuando ésta se repite varias veces. En este sentido, el término fiabilidad se usa habitualmente como sinónimo de repetibilidad, reproducibilidad o concordancia. Posteriormente, sin embargo, se expondrá que el concepto de fiabilidad es más sutil de lo que estos términos sugieren.

El término validez se refiere a si el procedimiento está midiendo realmente el fenómeno que queremos medir. La reproducibilidad del instrumento de medida es un prerrequisito de su validez, de manera que antes de plantearnos si el instrumento mide lo que queremos medir, debemos asegurarnos de que el instrumento mide «algo» de una forma reproducible: si el instrumento de medida no ofrece resultados reproducibles, entonces el instrumento de medida no es fiable, y resulta ocioso plantearnos el problema de la validez.

CONCEPTO DE FIABILIDAD

El concepto de fiabilidad y los diversos índices usados para estimarla se comprenden mejor si se hace explícito el modelo de medida utilizado. Para una variable aleatoria el modelo más sencillo es:

$$X = Y + \varepsilon \qquad [1]$$

donde X, Y y ε son variables que representan, respectivamente, el resultado de la medición, la magnitud a medir y el error de la medición. Este modelo también se puede escribir como:

$$X = \mu + \varepsilon_Y + \varepsilon \qquad [2]$$

donde se ha descompuesto la variable Y en la suma de su media μ (constante) y la variable ε_Y que contiene toda la variabilidad de Y alrededor de la media y cuya media, por lo tanto, será cero. Aunque tanto ε_Y como ε son variables aleatorias, la primera representa la variabilidad de la variable a medir (variabilidad biológica), mientras que la segunda representa el error de la medición. Hay que hacer notar, además, que lo único observable en [2] es X.

Si en [2] se calculan el valor esperado y la varianza, suponiendo independencia entre ε_Y y ε, se obtiene:

$$E(X) = \mu + E(\varepsilon) \qquad Var(X) = Var(\varepsilon_Y) + Var(\varepsilon)$$

A E (ε) se le denomina *error sistemático* o *sesgo*. La Var (ε) nos indica el error aleatorio (debido a la necesidad de conservar las unidades, el error aleatorio se define como la desviación típica de ε). Es importante notar que no se trata tanto de dos errores distintos como de dos aspectos distintos (e independientes si se acepta, como es habitual, la distribución gaussiana) del error de la medición. De la primera expresión se deduce que una medida de la validez es E (ε) = E (X) – μ; y de la segunda, que la variancia de la medición tiene dos componentes: una es la variancia de la propia variable y otra la del instrumento de medida. La variancia de la medición no es, pues, un buen indicador de la estabilidad de la medición; sí lo sería, sin embargo, si la magnitud a medir fuera constante, es decir, si la variancia de ε_Y fuera cero, como sería el caso si se tratara de estimar la precisión de una balanza usando unas pesas estándar. El coeficiente de reproducibilidad[6] se define como:

$$\rho = \frac{Var(\varepsilon_Y)}{Var(\varepsilon_Y) + Var(\varepsilon)} \qquad [3]$$

es decir, como la proporción de la variabilidad de la medición que se debe realmente a las variaciones de la variable, típicamente variaciones entre individuos.

En el modelo [3] se está considerando como el «instrumento» a un todo que incluye tanto al instrumento de medida propiamente dicho como al observador que lo utiliza. Este modelo es útil cuando el instrumento es muy simple o muy automático. Sin embargo, a veces en clínica es útil añadir en el modelo otra variable para la variabilidad introducida por el observador (ε_o), que da lugar a otro componente de la variancia de la medición. La fórmula [3], asumiendo también que esta nueva variable es independiente de las otras se modifica:

$$\rho = \frac{\text{Var}(\varepsilon_Y)}{\text{Var}(\varepsilon_Y) + \text{Var}(\varepsilon_o) + \text{Var}(\varepsilon)} \quad [4]$$

En el ámbito de las ciencias sociales a este cociente de la variancia debida a los individuos y a la variancia total se le denomina formalmente fiabilidad (*reliability*) y nos indica en qué grado el instrumento de medida es capaz de diferenciar entre individuos.

Aunque en un principio puede resultar algo extravagante, la idea de fiabilidad como proporción de la variancia total no es ajena a la práctica clínica habitual. Supongamos, por ejemplo, que en un paciente clínicamente estable, determinamos la presión intraarterial sistólica sistémica mediante dos transductores de presión, y obtenemos una diferencia entre las dos mediciones de 5 mmHg. La mayoría de los lectores considerarían esta diferencia como aceptable. Esta misma diferencia resultaría probablemente excesiva tratándose de una medición de la presión arterial pulmonar. ¿Cuál es la razón de que la misma diferencia sea aceptable en un caso e inaceptable en el otro? Probablemente, ello se debe a que lo que cuenta para aceptar el instrumento como fiable no es la magnitud del error sino la relación entre éste y el rango en que se mueven habitualmente las mediciones (que suele ser desconocido en los índices de nueva creación)[7].

De estas definiciones se desprende que la fiabilidad aumenta si se reduce el error de medida (ε), pero también si aumenta la variabilidad entre sujetos (ε_Y). Por lo tanto, la fiabilidad no es una propiedad intrínseca del instrumento, sino que depende de la variabilidad de la población a la que se aplique.

La estimación de la fiabilidad se hace básicamente mediante un análisis de los componentes de la variancia, determinando qué parte de la variabilidad es atribuible a los sujetos examinados (pacientes), a los observadores (en el modelo [4]) y al error del método o instrumento. En la práctica, sin embargo, la medición de la fiabilidad se aborda de diversas maneras, dependiendo del tipo de variables y del diseño del estudio.

La terminología usada en medición está llena de sinónimos, polisemias y términos confusos[8]. El término «fiabilidad», por ejemplo, puede resultar engañoso, ya que suscita en el lector la idea de confianza en el resultado, algo que no depende en exclusiva de la fiabilidad: una báscula sesgada puede proporcionar mediciones absolutamente idénticas sin que por ello el resultado merezca nuestra confianza. Algunos autores[1] proponen la sustitución del término fiabilidad por consistencia (*consistency)*. El problema de este término es que, aunque refleja de manera adecuada el significado habitual de fiabilidad como sinónimo de reproducibilidad o concordancia, no abarca globalmente el concepto clásico de fiabilidad.

En este artículo usaremos de forma preferente en lo sucesivo el término concordancia para referirnos a la estabilidad de la medición, y reservaremos el término fiabilidad para referirnos al concepto clásico, en el sentido de capacidad de distinguir entre individuos. El lector puede encontrar un glosario de términos al final de este artículo.

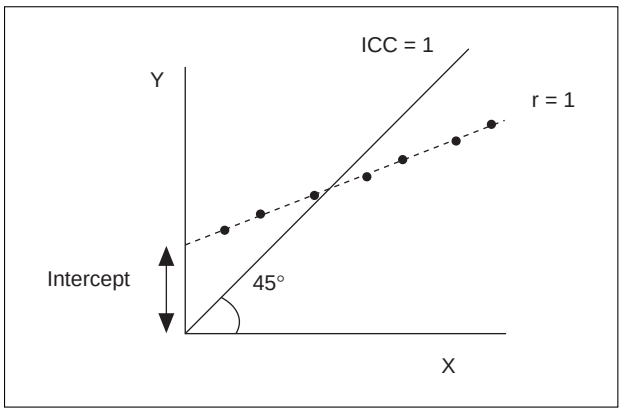

Fig. 1. Ejemplo de mala concordancia con r próxima a 1.

CUANTIFICACIÓN DE LA CONCORDANCIA: VARIABLES CONTINUAS

Coeficiente de correlación (r de Pearson)

Cuando la variable utilizada es continua y existen dos observaciones (instrumentos u observadores) por sujeto resulta intuitivo representar cada par de observaciones en un diagrama de puntos, examinar gráficamente si existe relación lineal entre las dos variables y calcular el coeficiente de correlación (r). Se esperaría que cuanto más coincidan las mediciones, más se aproximará a una recta el diagrama de puntos.

Aunque este uso de la r está muy difundido en la bibliografía médica, la r no mide realmente concordancia sino sólo el grado en que los puntos del diagrama de dispersión se ajustan a una recta. El uso de la r como índice de concordancia presenta varios problemas[9,10]:

1. Puede haber una r = 1 con una mala concordancia. Ello ocurre cuando la constante es distinta de cero, o el coeficiente de regresión distinto de uno: la relación entre las dos variables es lineal, pero las dos medidas no son equivalentes (fig. 1).

2. La r depende de la variabilidad entre los sujetos: si se incluyen valores extremos (como se suele hacer en los estudios de medición) el coeficiente de correlación aumenta. Por lo tanto, el valor de r publicado en un artículo sólo se puede extrapolar a nuestra población cuando ésta es parecida a aquella en la que se ha calculado.

Ejemplo: la tabla 1 representa los datos de un estudio hipotético en el que 20 ecocardiogramas fueron

TABLA 1
Resultados de la evaluación de la fracción de eyección por ecocardiograma en 20 pacientes (2 observadores)

Paciente	Observador 1	Observador 2	Media	Diferencia
1	0,27	0,32	0,30	–0,05
2	0,30	0,35	0,33	–0,05
3	0,30	0,32	0,31	–0,02
4	0,32	0,32	0,32	0,00
5	0,35	0,40	0,38	–0,05
6	0,35	0,32	0,34	0,03
7	0,37	0,40	0,39	–0,03
8	0,37	0,35	0,36	0,02
9	0,37	0,45	0,41	–0,08
10	0,38	0,35	0,37	0,03
11	0,38	0,42	0,40	–0,04
12	0,39	0,40	0,40	–0,01
13	0,40	0,38	0,39	0,02
14	0,40	0,42	0,41	–0,02
15	0,42	0,45	0,44	–0,03
16	0,43	0,40	0,42	0,03
17	0,45	0,50	0,48	–0,05
18	0,48	0,45	0,47	0,03
19	0,52	0,60	0,56	–0,08
20	0,55	0,50	0,53	0,05

evaluados por dos observadores independientes, que estimaron la fracción de eyección. Si ordenamos los datos en función de la magnitud de la fracción de eyección estimada por el observador 1 y dividimos el conjunto en 2 bloques de 10 pacientes, la r de cada uno de los bloques (0,57 y 0,78, respectivamente) es menor que la r del conjunto de los pacientes que incluye los valores extremos (0,85).

3. La r se calcula a partir de los pares (x,y) ordenados de mediciones; si en algunas observaciones se cambiara el orden, el valor de r cambia, lo que no es deseable en un índice de repetibilidad, porque los cambios que tengan que ver con el orden de las mediciones están indicando diferencias entre los valores medios.

Ejemplo: si en la tabla 1 se invierten los valores de las parejas impares la r pasa de 0,85 a 0,91.

El coeficiente de correlación intraclase

Es la estimación del coeficiente [3] o [4], y se representa habitualmente como r_I o CCI (ICC en inglés). Constituye un mejor índice que la r de Pearson como medida de fiabilidad. El r_I estima la correlación promediada entre todas las posibles ordenaciones de los pares de observaciones.

Tiene dos ventajas claras: en primer lugar, se obvia el problema de la dependencia del orden. En segundo lugar, permite extender el cálculo a la situación de más de dos observaciones por sujeto[7,9]. Su estimación, sin embargo, plantea alguna dificultad por las distintas formas que adopta dependiendo del diseño del estudio[11]. Una de las fórmulas de cálculo más habituales se basa en la tabla de ANOVA:

$$r_I = \frac{mSS_B - SS_T}{(m-1)\,SS_T} \qquad [5]$$

donde m es el número de observaciones por sujeto, SS_B es la suma de cuadrados entre los sujetos y SS_T la suma de cuadrados total.

Ejemplo: para los datos de la tabla 1,

$$r_I = \frac{2 \times 0{,}2184 - 0{,}201}{(2-1) \times 0{,}201} = 0{,}83$$

y este resultado no cambia aunque se invierta el orden de las observaciones impares.

En la práctica, suele haber poca diferencia entre r y r_I. El r_I puede ser mucho menor que r cuando existe un cambio sistemático entre la primera y la segunda mediciones (p. ej., cuando existe efecto de aprendizaje). No obstante, en este caso no se darían las condiciones para un verdadero análisis de fiabilidad, se estaría introduciendo un error sistemático y carecería de sentido plantearse el problema de la fiabilidad.

Método de Bland-Altman

Ha alcanzado una gran difusión para analizar la concordancia entre 2 métodos que utilizan las mismas unidades de medida. Consiste en representar gráficamente la diferencia entre las dos observaciones contra su media (fig. 2B). Ello permite examinar rápidamente la magnitud de las discrepancias y su relación con la magnitud de la medición. Adicionalmente se puede estimar el error estándar de las diferencias y las bandas de confianza entre las que cabe esperar que se encuentre el 95% de las diferencias.

El procedimiento de cálculo es sencillo, siempre que la variabilidad sea constante. Cuando ésta depende de la magnitud de la medición (p. ej., cuando la variabilidad es mayor conforme aumenta la magnitud de la medida), el cálculo se complica ligeramente[12]. Por ello es siempre aconsejable representar gráficamente la desviación estándar de las medidas repetidas para cada media.

Tiene la ventaja sobre la r de Pearson de que informa explícitamente de la magnitud de las discrepancias entre cada par de observaciones. Se ha argumentado que tiene la ventaja de que es independiente de la variación en la muestra de las observaciones. Sin embargo, este punto es objeto de discusión[7].

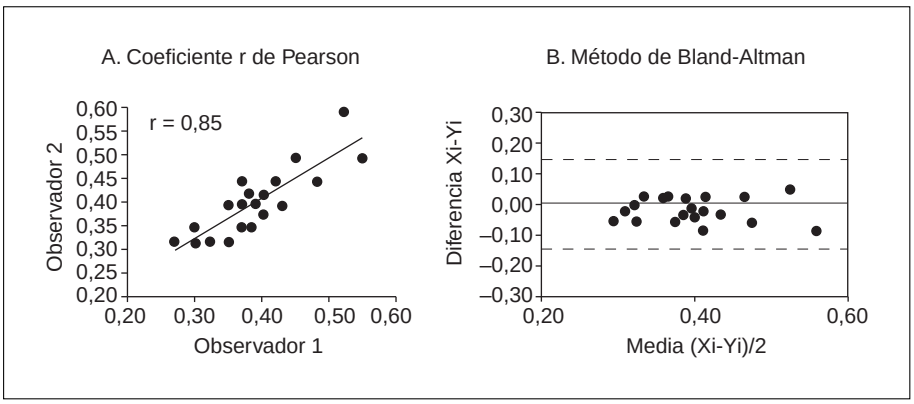

Fig. 2. Análisis gráfico de los datos de la tabla 1. A. regresión lineal; B: método de Bland y Altman.

CUANTIFICACIÓN DE LA CONCORDANCIA: VARIABLES DISCRETAS

Concordancia simple

La forma más sencilla de expresar la concordancia cuando la variable es categórica es a través de la proporción o porcentaje de observaciones concordantes (P_o), también llamado índice de acuerdo observado o concordancia simple. El problema es que una parte de esta concordancia se debe al azar.

Ejemplo: la tabla 2A representa los resultados hipotéticos de una prueba muy poco fiable, pero con una prevalencia de enfermos muy baja. En este caso la concordancia observada es del 88%, lo que puede sugerir al lector poco avezado que se trata de una prueba muy fiable. Resulta obvio, sin embargo, que gran parte de esta variabilidad se debe sencillamente al azar. Interesa, por lo tanto, cuantificar el grado de concordancia más allá del azar. Uno de los métodos más utilizados para medirla es el índice kappa.

Índice kappa

El grado de concordancia esperable por azar se puede calcular a partir del producto de los marginales de la tabla de contingencia.

Ejemplo: la tabla 2B presenta los valores esperados para cada una de las celdas de la tabla 2A, así como el procedimiento de cálculo.

El coeficiente kappa, que se calcula mediante la siguiente expresión:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad [6]$$

donde P_o es la proporción de concordancia observada y P_e es la proporción esperada.

La conceptualización del índice kappa se facilita mediante su representación gráfica[5,13]. La figura 3 representa los datos obtenidos del problema enunciado en la tabla 2. En la figura se puede ver que la mayoría de la concordancia observada es atribuible al azar: lo que interesa es medir el grado de concordancia no atribuible al azar (el área entre 0,84 y 1,00). El coeficiente kappa representa el 27% de dicha área.

TABLA 2
Concordancia en una tabla 2 × 2

2A: valores observados

		Observador A		
		Enfermo	Sano	Total
Observador B	Enfermo	a = 3	b = 6	fl = 9
	Sano	c = 6	d = 85	f2 = 91
	Total	cl = 9	c2 = 91	t = 100

2B: valores esperados

		Observador A		
		Enfermo	Sano	Total
Observador B	Enfermo	c1 × f1/t = 0,81	c2 × f1/t = 8,19	f1 = 9
	Sano	c1 × f2/t = 8,19	c2 × f2/t = 82,81	f2 = 91
	Total	cl = 9	c2 = 91	t = 100

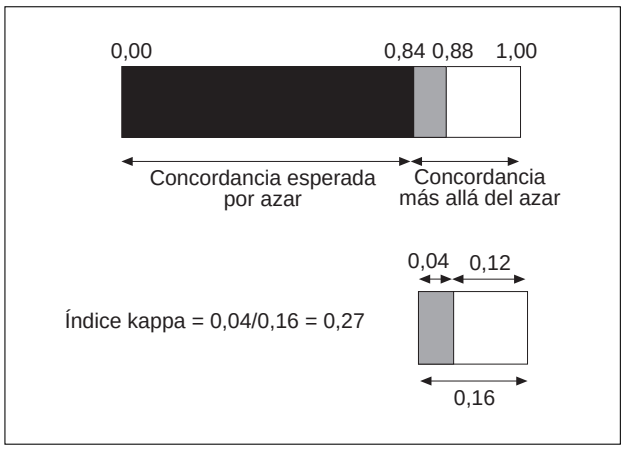

Fig. 3. Conceptualización del índice kappa para los datos de la tabla 2.

El valor de kappa es de 1 si hay total acuerdo; si el acuerdo observado es igual al esperado, kappa vale 0, y es menor de 0 si el acuerdo observado es menor que el esperado por azar. En general, se aceptan los niveles descritos en la tabla 3.

La misma expresión [6] permite calcular el índice kappa cuando existen más de 2 categorías nominales. La tabla 4 ofrece un ejemplo de cálculo de concordancia entre 2 observadores, para 3 categorías.

Existen extensiones del índice kappa que permiten abordar diversos problemas. Aquí enumeraremos sólo algunas de ellas; el lector interesado puede remitirse a la bibliografía seleccionada, en especial, al libro de Fleiss[14], en el que se pueden encontrar, asimismo, las fórmulas para el cálculo del error estándar de kappa y de su intervalo de confianza al 95%. He aquí alguna de estas extensiones:

TABLA 3
Interpretación de los valores de 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

Índice kappa específico

En ocasiones, además del kappa global, interesa conocer la concordancia para una categoría específica.

Ejemplo: en un estudio[15] encaminado a evaluar la fiabilidad del diagnóstico radiológico de insuficiencia cardíaca a partir de las radiografías de tórax obtenidas en la unidad coronaria, se le pidió a 4 radiólogos que examinaran cada uno de ellos 25 radiografías de pacientes ingresados en la unidad coronaria por infarto agudo de miocardio, y que las clasificaran en 4 grupos: insuficiencia cardíaca severa, ligera moderada, ausente o placa no valorable. La concordancia (kappa) global fue de 0,34. Sin embargo, esta concordancia global ocultaba una concordancia específica aceptable para las categorías insuficiencia cardíaca severa (kappa = 0,50) o ausencia de insuficiencia car-

TABLA 4
Concordancia para 3 categorías

4A: valores observados

		Observador A			
		Enfermo	Dudoso	Sano	Total
Observador B	Enfermo	3	3	3	9
	Dudoso	3	10	5	18
	Sano	3	5	65	73
	Total	9	18	73	100

4B: valores esperados

		Observador A			
		Enfermo	Dudoso	Sano	Total
Observador B	Enfermo	0,81	1,62	6,57	9
	Dudoso	1,62	3,24	13,14	18
	Sano	6,57	13,14	53,29	73
	Total	9	18	73	100

Concordancia observada = (3 + 10 + 65)/100 = 0,78.
Concordancia esperada = (0,81 + 3,24 + 53,29)/100 = 0,5734.
Kappa = (0,78-0,5734)/(1-0,5734) = 0,2066/0,4266 = 0,48.

TABLA 5
Ejemplo de cálculo del índice kappa ponderado (pesos cuadráticos); 2 observadores, 4 categorías ordinales. En negrita se presentan los valores observados; en el vértice superior derecho de cada casilla, los valores esperados y en el vértice inferior derecho los pesos cuadráticos

	A	B	C	D	
A	2,64 **6** 0	2,64 **3** 1	2,64 **2** 4	3,08 **0** 9	**11**
B	3,36 **4** 1	3,36 **7** 0	3,36 **2** 1	3,92 **1** 4	**14**
C	1,44 **1** 4	1,44 **0** 1	1,44 **3** 0	1,68 **2** 1	**6**
D	4,56 **1** 9	4,56 **2** 4	4,56 **4** 1	5,32 **11** 0	**19**
	12	**12**	**12**	**14**	**50**

$\Sigma w_{ij}P_o$ = [1 (4 + 0 + 5 + 3 + 2 + 2) + 2 (1 + 2 + 2 + 1) + 4 (1 + 0)]/50 = 0,98.
$\Sigma w_{ij}P_e$ = [1 (3,36 + 1,44 + 4,56 + 2,64 + 3,36 + 1,68) + 4 (1,44 + 4,56...)]/50 = 2,73.
Kappa = 1 – 0,98/2,73 = 0,64.

díaca (kappa = 0,45), y una concordancia inaceptable para las categorías insuficiencia cardíaca ligera/moderada (kappa = 0,19) o placa no interpretable (kappa = 0,15).

Índice kappa ponderado

El índice kappa, en cualquiera de sus formulaciones anteriores, sólo tiene en cuenta la concordancia exacta. En ocasiones, sin embargo, unas formas de discordancia son más graves que otras. Por ejemplo, un par de observaciones del tipo enfermo/sano es más grave que un par de observaciones del tipo enfermo/dudoso. El llamado índice kappa ponderado tiene en cuenta este acuerdo aproximado; para ello, las celdillas de acuerdo total (situadas en la diagonal que va desde arriba a la izquierda hasta abajo a la derecha) tienen un peso de 0, mientras que las celdillas de los ángulos opuestos tienen la máxima ponderación. En su formulación habitual, el coeficiente kappa ponderado se calcula mediante la expresión:

$$\kappa_w = 1{,}0 - \frac{\Sigma w_{ij} \times Po_{ij}}{\Sigma w_{ij} \times Pe_{ij}}$$

donde κ_w es el coeficiente kappa ponderado, w_{ij} es el peso asignado a la celda i,j, y Po_{ij} y Pe_{ij} son, respectivamente, las proporciones observadas y esperadas de la celda i,j.

En principio, el peso otorgado a cada grado de discrepancia es arbitrario, lo que dificulta su interpretación[16]; sin embargo, lo más habitual es usar como esquema de ponderación los pesos cuadráticos (basados en el cuadrado de la discrepancia). Cuando se utiliza este esquema, el kappa ponderado coincide aproximadamente con el coeficiente de correlación intraclase (y exactamente cuando las distribuciones marginales son idénticas)[7]. En la tabla 5 se presenta un ejemplo de cálculo.

Índice kappa con múltiples observadores

Las identidades entre el coeficiente kappa y el coeficiente de correlación intraclase han sido explotadas algebraicamente para obtener un índice kappa aplicable a la situación en que cada individuo es observado un número de veces m.

PROBLEMAS DEL ÍNDICE KAPPA

Sin duda, el índice kappa supone una mejora con respecto al porcentaje de concordancia observada. Sin embargo, en los últimos años han aparecido una gran cantidad de artículos que señalan la existencia de limitaciones del índice kappa y sugieren métodos alternativos para cuantificar la concordancia en variables categóricas.

En el caso de variables nominales, el coeficiente kappa valora igual una discrepancia severa que una discrepancia despreciable. Obviamente, cuanto mayor es el número de categorías, menor es la probabilidad de obtener una concordancia «exacta». En consecuencia el coeficiente kappa depende sensiblemente del número de categorías, de manera que disminuye conforme aumenta su número. Además, el coeficiente kappa varía dependiendo más de la forma en que fueron elegidas las categorías que del grado de reproducibilidad de los métodos. Por ello, cuando haya más de 2 categorías puede ser conveniente comparar cada una de ellas (concordancia específica) con la suma de las demás.

El coeficiente kappa ponderado, aunque resuelve en principio el problema del acuerdo exacto, plantea un nuevo problema: el de la elección de los pesos. Si los pesos se eligen arbitrariamente, se dificulta la comparación entre estudios; por el contrario, si la ponderación se hace por pesos cuadráticos, entonces el kappa es equivalente al coeficiente de correlación intraclase, que es menos sensible a los cambios en el número de categorías y tiende a aumentar más que a disminuir con el número[16].

En segundo lugar, el coeficiente kappa depende de la prevalencia de las categorías. Cuando una categoría presenta una prevalencia muy alta o muy baja, aun manteniendo constante la calidad de la medición, el índice kappa disminuye. Ello hace que dos índices

kappa brutos, obtenidos en poblaciones con distinta prevalencia del rasgo en estudio, no puedan ser comparados entre sí.

Con el objeto de obviar este problema, se ha propuesto[17] el uso del *kappa máximo*, es decir, el mayor de los kappa obtenidos para las diversas prevalencias de la característica medida, manteniendo fijas la sensibilidad y especificidad de las dos mediciones. El problema es que habitualmente no se conoce la sensibilidad/especificidad del instrumento, por lo que sus valores sólo pueden asumirse.

En los últimos años ha proliferado la bibliografía sobre el índice kappa. Feinstein y Cichetti[18,19] han subrayado el efecto «indeseable» sobre el índice kappa del balance y la simetría entre los marginales totales, y proponen las correspondientes medidas correctoras. Diversos autores[20,21] han propuesto la modelización de la concordancia mediante modelos estadísticos como alternativa al kappa. Otros autores[22,23] han discutido el papel del índice kappa como índice de validez. El lector interesado en profundizar en estos aspectos puede consultar la bibliografía referenciada al final del artículo.

FORMAS DE EVALUAR LA CONCORDANCIA. CONSISTENCIA INTERNA

El diseño del estudio de medición condiciona las distintas formas de evaluar la fiabilidad/concordancia del instrumento de medida. En el ámbito de los índices clínicos se utilizan habitualmente algunas de las siguientes estrategias[7]:

1. Cuando el instrumento de medida esté pensado para que los datos sean recogidos por varios observadores (no necesariamente siempre los mismos) es esencial cuantificar la variabilidad entre los distintos observadores. Para ello, se pide a varios observadores que examinen al mismo sujeto. La concordancia medida de esta forma se denomina *concordancia interobservadores*.

2. Si el instrumento de medida va a ser utilizado por un solo observador en varios puntos en el tiempo interesa cuantificar la *concordancia intraobservador*, que nos ofrece una idea de lo repetitivo de los resultados.

3. En el caso de cuestionarios autoadministrados, la concordancia se suele medir pasando al individuo el mismo cuestionario en dos ocasiones distintas. Cuando la concordancia se determina de este modo se denomina *concordancia test-retest.*

En el caso de índices elaborados a partir de varios ítems, además de la concordancia, entendida como estabilidad de los resultados tras varias administraciones de la prueba (concordancia interobservadores, intraobservadores o test-retest) es frecuente que se evalúe la *consistencia interna* de la escala en una sola administración. La idea subyacente –extraída de la psicometría– presupone que si varios ítems están midiendo el mismo rasgo, estos ítems deberían estar intercorrelacionados. Este grado de intercorrelación se puede medir de varias maneras, siendo el coeficiente alfa de Cronbach uno de los estadísticos más utilizados[1,7], y es frecuente que se haga referencia a este estadístico como la «fiabilidad del test».

Obviamente, la «fiabilidad» medida por la consistencia interna del índice ignora algunos factores que hacen que la medida varíe de un observador a otro o de una medición a otra posterior, por lo que tiende a ser mayor que la fiabilidad medida a partir de varias mediciones. Por lo general, se exige que la consistencia interna (alfa de Cronbach) sea superior a 0,8, mientras que la fiabilidad mínima exigida cuando se mide por criterios externos oscila alrededor de 0,55.

La utilización del índice de Cronbach (y los demás estadísticos relacionados) en clínica debe hacerse de forma juiciosa[1]. En efecto, mientras que en psicometría la intercorrelación entre los distintos ítems que están midiendo la misma dimensión suele ser un requisito básico, el objetivo básico en clinimetría no es la unidimensionalidad, por lo que puede ser en ocasiones aconsejable incluir dentro de la misma escala ítems no relacionados entre sí (p. ej., ítems relacionados con insuficiencia cardíaca e ítems relacionados con isquemia miocárdica). En estos casos, la existencia de un índice de Cronbach alto puede no sólo no ser exigible sino ni siquiera deseable.

VALIDEZ

En términos generales, la validez expresa la relación entre la medida y lo que queremos medir, así como con lo que no queremos medir[24]. En términos más técnicos, está relacionada con la cantidad de error sistemático (sesgo) introducido en la estimación. El sesgo es una amenaza para la validez y una medida sesgada puede llevarnos a conclusiones completamente erróneas. Por ejemplo, si tratamos de medir la presión arterial con un termómetro, comprobaremos que el resultado es reproducible, sin que ello signifique que está midiendo realmente la presión arterial.

Evaluar la validez de un instrumento exigiría comprobar sus resultados con el verdadero valor que se trata de medir, es decir, disponer de otro instrumento (patrón de oro o *gold standard*) que permita saber ese verdadero valor. Esto, en la clínica, plantea diversos problemas:

1. El patrón puede existir, pero ser lo suficientemente agresivo como para no poder usarlo, sin problemas éticos, en un estudio de validez. Por ejemplo, el diagnóstico de infarto no se hace normalmente recurriendo al *gold standard* (la anatomía patológica) sino a una prueba diagnóstica rápida, inocua (e imperfecta) como es el

electrocardiograma de superficie. En algunos casos se podría disponer del resultado del *gold standard* (p. ej., pacientes en los que es necesaria la biopsia), pero esto generaría una muestra sesgada: se pueden estimar bien los falsos positivos, pero no los falsos negativos.

2. También existen dificultades para evaluar la validez de un nuevo instrumento (p. ej., la ecocardiografía intraesofágica) al que se supone mejor que las pruebas de referencia anteriores (p. ej., el ecocardiograma convencional o la angiografía).

3. Por último, resulta, asimismo, problemático evaluar la validez de índices complejos construidos para cuantificar constructos como calidad de vida o gravedad de la insuficiencia cardíaca, en los que no existe un patrón de referencia claro.

En otro artículo de esta serie se pondrá énfasis en las dos primeras situaciones; en este artículo nos limitaremos fundamentalmente a este último problema.

Clásicamente la validez se ha estudiado mediante las «tres C» (Contenido, Criterio y Constructo). Aunque en realidad los 3 enfoques están midiendo en esencia lo mismo –en qué medida la información que nos da el test nos permite inferir la magnitud del rasgo que queremos medir en un individuo[7]–, los trataremos separadamente por razones didácticas.

Validez de contenido

La validez de contenido indica hasta qué punto el conjunto de los ítems que forman el índice cubre las diferentes áreas o dominios que se quieren medir. La idea subyacente es que si el índice mide realmente el rasgo de interés debe incluir como ítems todas aquellas facetas que se consideran relevantes. Por ejemplo, si queremos utilizar un instrumento para medir la calidad de vida en enfermos coronarios parece exigible que dicho índice incluya ítems relacionados tanto con la capacidad física como con la presencia de ángor o de disnea.

En realidad, este aspecto de la validez tiene un límite temporal: a medida que se adquieren más conocimientos es probable que el instrumento necesite ser remodelado o planteado nuevamente en función de lo drástico que sea el cambio. Si pensamos en la tasa de morbilidad de individuos seropositivos para el VIH (virus de la inmunodeficiencia humana) en relación a las distintas definiciones que han ido surgiendo sobre el sida (síndrome de inmunodeficiencia adquirida), podemos comprender cómo la validez de contenido cambia paralelamente a estas nuevas definiciones.

Muy relacionada con la validez de contenido está la *validez aparente (face validity)* que examina si cada ítem incluido en el instrumento está claramente relacionado con el rasgo que deseamos medir. Por ejemplo, si queremos medir el grado de disnea, las preguntas sobre si al paciente le falta aire cuando sube escaleras o sobre si tiene que incorporarse en la cama durante las noches tienen validez aparente, mientras que la agudeza visual y la presencia de acúfenos no la tienen.

La validez de contenido y la validez aparente se determinan subjetivamente por parte de expertos. Los siguientes «tipos de validez», sin embargo, requieren evidencia objetiva para evaluarlos.

Validez de criterio

Se establece examinando la correlación entre el índice nuevo y una prueba de referencia (*gold standard*). La idea es que si los dos instrumentos están midiendo el mismo rasgo, sus resultados cuando se aplican a los mismos sujetos deben estar correlacionados.

Un ejemplo de validez de criterio lo constituye la evaluación del rendimiento de una prueba diagnóstica mediante una tabla 2×2, en la que el resultado (positivo o negativo) de aplicar la prueba a un grupo de sujetos se compara con el veredicto de la prueba de referencia, lo que nos indica el número de sujetos bien clasificados (como enfermos o sanos), el número de falsos positivos y el número de falsos negativos.

Básicamente la validez de criterio se puede evaluar de dos formas:

1. Cuando los dos instrumentos de medida se pasan de forma más o menos simultánea (p. ej., cuando se crea un nuevo instrumento de medida) se dice que se está evaluando la validez concurrente.

2. Otras veces, el patrón de referencia no está disponible inmediatamente, por ejemplo cuando el criterio se basa en la evolución a largo plazo (comprobación por autopsia). En este caso, se dice que se está evaluando la validez predictiva.

En el caso de la evaluación de las pruebas diagnósticas, la validez de criterio se examina de forma peculiar, separando dos parámetros diferentes: la sensibilidad (proporción de casos que dan positivo a la prueba) y la especificidad (proporción de no casos que dan negativo a la prueba-test) (véase el artículo correspondiente al diagnóstico, dentro de esta misma serie).

Validez de constructo (o de concepto)

Cuando no existe una prueba de referencia aceptable para evaluar la validez de la medición, se recurre con frecuencia a la evaluación de la validez de constructo. Un constructo es una especie de miniteoría, que explica y da coherencia a unos datos; ejemplos de constructo son la insuficiencia cardíaca, la ansiedad o la calidad de vida. La validez de constructo de un índice se evalúa poniéndolo a prueba mediante el diseño de pequeños experimentos. Ello se puede hacer de muchas maneras. Aquí veremos sólo dos de ellas.

1. Evaluación por grupos extremos. Se basa en aplicar el índice a pacientes a los que se supone (por crite-

rio de expertos) que poseen niveles extremos (máximos y mínimos) del rasgo que deseamos medir. Por ejemplo, para ver si un índice mide realmente capacidad funcional, examinaríamos los índices obtenidos en un grupo de personas muy incapacitadas y en un grupo de sujetos activos normales y compararíamos los resultados. Habitualmente, sin embargo, nuestro interés no consiste en identificar grupos extremos sino en distinguir entre sujetos que tienen niveles intermedios del rasgo. Por ello, si nos limitamos a hacer este tipo de evaluación podemos sobreestimar la capacidad del instrumento.

2. Validez convergente/discriminante: si el índice es válido, debería correlacionarse con otros índices que miden la variable de interés (*validez convergente*), mientras que no debería mostrar ninguna correlación con otros índices que miden variables distintas (*validez discriminante*). Por ejemplo, supongamos que estamos evaluando un índice electrocardiográfico de tamaño de infarto; si realmente el índice mide tamaño de infarto, debe existir una correlación entre éste y el pico sérico de CK durante la fase aguda o la fracción de eyección por isótopos (validez convergente); el índice no debe mostrar una correlación, sin embargo, con variables poco o nada relacionadas con el tamaño de infarto como la edad o el valor del colesterol.

CONSTRUCCIÓN Y ADAPTACIÓN DE ESCALAS DE SALUD

Como vemos, por tanto, la tarea de probar la fiabilidad y validez de una escala de salud («validar la escala») es una tarea compleja. El proceso de validación involucra numerosas fuentes de información y la recogida de evidencias empíricas que apoyan la validez de la medida es un proceso interminable. Es de resaltar que no hay un criterio concreto a partir del cual la medida se considere válida. Se puede afirmar, por tanto, que un instrumento de medida de salud no puede ser totalmente validado en un solo estudio. También es de destacar que, si una escala de salud es válida para un propósito, no es necesariamente válida para otro[25]. Hoy día existen buenos repertorios sobre escalas de salud[26], que se deberían consultar antes de embarcarse en la tarea de elaborar un índice nuevo.

Un problema peculiar es el de la necesidad de adaptar cuestionarios escritos en otros idiomas. En efecto, la traducción de los ítems puede generar distorsiones sutiles en su significado, hasta el punto de que algunos han llegado a afirmar que un cuestionario traducido es en realidad un «nuevo cuestionario». Por ello, la traducción del instrumento ha de llevarse a cabo de una manera no tradicional. El proceso consiste en una serie de traducciones que van desde la lengua original a la nueva y viceversa (retrotraducciones), tantas veces como sea necesario, hasta conseguir un instrumento que culturalmente sea compatible con el instrumento original. Se trata de conservar así un alto grado de validez aparente del nuevo instrumento[27]. De esta manera se consigue que la importación de instrumentos de medida sea ventajosa y se evita que el proceso de validación empiece desde cero. La validez del instrumento de medida no se mide, sin embargo, en una escala de todo o nada, sino que tiene una gradación, lo que obliga a sopesar en cada caso concreto el tipo de adaptación que resulta adecuado.

BIBLIOGRAFÍA

1. Feinstein AR. Clinimetrics. New Haven: Yale University Press, 1987.
2. Feinstein AR. A bibliography of publications on observer variability. J Chron Dis 1985; 38: 619-632.
3. Sackett DL. A primer on the precision and accuracy of the clinical examination. JAMA 1992; 267: 2.638-2.644.
4. Department of Clinical Epidemiology and Biostatistics, McMaster Univerity. Clinical Disagreement: I. How often it occurs and why. CMA Journal 1980; 123: 499-504.
5. Hernández Aguado I, Porta Serra M, Miralles M, García Benavides F, Bolúmar F. La cuantificación de la variabilidad en las observaciones clínicas. Med Clin (Barc) 1990; 95: 424-429.
6. Fisher RA. Statistical methods for research workers. Nueva York: Hafner, 1958.
7. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use (2.ª ed.). Oxford: Oxford University Press, 1995.
8. Last JM, editor. A dictionary of epidemiology (3.ª ed.). Nueva York: Oxford University Press, 1995.
9. Bland JM, Altman DG. Measurement error and correlation coefficients. BMJ 1996; 313: 41-42.
10. Bland JM, Altman DG. Measurement error. BMJ 1996; 312: 1.654.
11. Bartko JJ. The intraclass correlation coeficient as a measure of reliability. Psychol Rep 1966; 19: 3-11.
12. Bland JM, Altman DG. Measurement error proportional to mean. BMJ 1996; 313: 106-107.
13. Sackett DL, Haynes RB, Tugwell P. Epidemiología Clínica. Una ciencia básica para la medicina clínica. Madrid: Díaz de Santos, 1989.
14. Fleiss JL. The measurement of interrater agreement. En: Fleiss JL, editor. Statistical methods for rates and proportions (2.ª ed.). Toronto: Willey, 1981; 212-236.
15. Arráez V, Latour J, García Benavides F, Giner S, Díaz Castellanos MA, Sánchez Candel F. Interpretación de la radiología torácica en pacientes de cuidados intensivos. Análisis de la variabilidad interobservadores [resumen]. Med Intensiva 1988; 2 (Supl 1): 287-288.
16. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987; 126: 161-169.
17. Walter SD. Measuring the reliability of clinical data: the case for using three observers. Rev Epidemiol Sante Publique 1984; 32: 206-211.
18. Feinstein AR, Cichetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43: 543-549.
19. Cichetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990; 43: 543-549.
20. May SM. Modelling observer agreement - an alternative to kappa. J Clin Epidemiol 1994; 47: 1.315-1.324.
21. Coughlin SS, Pickle LW, Goodman MT, Wilkens LR. The logistic modeling of interobserver variation. J Clin Epidemiol 1992; 45: 1.237-1.241.
22. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol 1988; 41: 949-958.

23. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. J Clin Epidemiol 1988; 41: 959-968.
24. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Nueva York: Van Noostrand Reinhold Company Inc., 1982; 183-267.
25. Ware JE, Brook RH, Davies AR, Lohrd K. Choosing measures of health status for individuals in general population. Am J Public Health 1981; 71: 620-625.
26. McDowell I, Newell C. Measuring health. Oxford: Oxford University Press, 1987.
27. Badia X, Alonso J. Validity and reproducibility of the spanish versionof the Sickness Impact Profile. J Clin Epidemiol 1996; 49: 359-365.

GLOSARIO DE TÉRMINOS USADOS EN ESTUDIOS DE MEDICIÓN

Medición: proceso mediante el cual se cuantifica una magnitud.

Instrumento de medida: habitualmente se utiliza en sentido amplio, incluyendo no sólo al aparato de medida en sentido estricto sino también al observador que lo utiliza. Se llaman instrumentos a un cuestionario sobre calidad de vida, a un ecocardiógrafo, a un cardiólogo examinando las ondas del pulso venoso yugular, o al conjunto formado por un equipo radiológico y el radiólogo que interpreta las imágenes, etc.

Validez (*validity*): cualidad del instrumento que mide la magnitud *real*. Como la magnitud real es siempre inaccesible, la validez no se puede calcular, sólo estimar o acotar. No existen instrumentos perfectamente válidos, de modo que se trata siempre de una cuestión de grado.

Calibración: proceso para controlar la validez de un instrumento. Exige comparar mediciones con un patrón de referencia. En este sentido se distingue entre exactitud y conformidad.

Exactitud (*accuracy*): grado en que la medición coincide con un patrón de referencia claro y objetivo.

Conformidad (*conformity*): grado en que la medición coincide con otra medida considerada como mejor pero no se puede considerar como un *gold standard*.

Fiabilidad (*reliability*): clásicamente se refiere a la capacidad del instrumento para distinguir entre individuos, independientemente de que esa medición sea o no válida. Se mide como la proporción de la variabilidad total que se debe a diferencias reales entre los sujetos, es decir, la proporción de la variabilidad que no se debe a los observadores o al error. En el área de la clínica y la epidemiología se utiliza habitualmente como sinónimo de repetibilidad o reproducibilidad. Aunque el término fiable sugiere a primera vista que el observador puede confiar en el resultado, no se debe confundir fiabilidad con validez: un instrumento de medida puede proporcionar resultados fiables pero sesgados y por tanto no confiables.

Precisión (*precision),* **repetibilidad (***repeatibility),* **reproducibilidad (***reproducibility)* o **concordancia (***agreement, concordance*): cualidad del instrumento que obtiene los mismos resultados al medir la misma magnitud. Validez y reproducibilidad son dos cualidades conceptualmente independientes, aunque como para estimarlas hay que realizar mediciones, son instrumentalmente dependientes. Además, si un instrumento no es aceptablemente preciso, no tiene sentido plantearse su validez.

Consistencia (*consistency*): sinónimo de concordancia. En epidemiología se utiliza con mucha frecuencia para indicar que una asociación se encuentra uniformemente en distintos grupos de individuos o en diferentes tipos de estudio.

Consistencia interna (*internal consistency*): indica en qué medida los ítems incluidos en una escala están interelacionados. Se mide habitualmente por el coeficiente alfa de Cronbach. Algunas veces se presenta el índice de Cronbach como la fiabilidad del test; esta interpretación debe ser tomada con precaución, ya que la consistencia interna se mide en una sola determinación y no tiene en cuenta otras fuentes de variabilidad.

Error sistemático o sesgo de medición (*bias*): grado en que la medición se desvía de su valor real. Cuando la magnitud a medir es aleatoria (en clínica siempre) se estima con la diferencia entre la media de distintas medidas y el valor esperado de la magnitud.

Error aleatorio: error que atenta contra la reproducibilidad. Si la magnitud a medir es fija se estima a través de la variancia (o la desviación típica, para mantener las mismas unidades) de una serie de medidas de la misma magnitud. Sin embargo, cuando la magnitud es una variable aleatoria, la variancia de la medición, en el modelo más simple, tiene dos componentes (la variancia de la propia variable y la del instrumento).

Variabilidad entre sujetos: variabilidad «real», debida a diferencias entre los sujetos observados.

Variabilidad residual o variancia del error: parte de la variancia total no debida a diferencias entre los sujetos.

Validez de contenido (*content validity*): define hasta qué punto la selección de ítems cubre las diferentes áreas o dominios que se quieren medir.

Validez aparente (*face validity*): examina si, a juicio de un experto, los ítems incluidos en una escala están relacionados con el rasgo que se desea medir.

Validez de criterio (*criterion validity*): se establece comparando el índice nuevo con una prueba de referencia (*gold standard*) y viendo cómo se correlacionan. La idea es que si los dos instrumentos están midiendo el mismo rasgo, sus resultados cuando se aplican a los mismos sujetos deben estar altamente correlacionados.

Validez de constructo (o de concepto) (*construct validity*): evalúa hasta qué punto la medida del instrumento en cuestión está correlacionado de otra medida de otro instrumento de una manera predictiva, pero para la cual no existe un verdadero criterio o patrón (*gold standard).*