C. Ochoa Sangrador[1], G. Orejas[2]

An Esp Pediatr 1999;50:301-314.

Epidemiología y metodología científica aplicada a la pediatría (IV): Pruebas diagnósticas

Introducción

El diagnóstico médico es un proceso dinámico en el que se intenta tomar decisiones idóneas en presencia de incertidumbre. En este proceso intervienen distintos instrumentos que tratan de reducir el grado de incertidumbre con el que se emiten los juicios diagnósticos. Junto a instrumentos clásicos, como la anamnesis y el examen físico, hoy en día disponemos de múltiples exploraciones complementarias que, usadas correctamente, permiten mejorar el proceso diagnóstico. Desde un punto de vista funcional, consideramos prueba diagnóstica a cualquier procedimiento realizado para confirmar o descartar un diagnóstico o incrementar o disminuir su verosimilitud. La utilidad de una prueba diagnóstica depende, fundamentalmente, de su validez y de su fiabilidad, pero también de su rendimiento clínico y de su coste.

El concepto de validez se refiere a la capacidad de la prueba para medir lo que realmente queremos medir. La validez se evalúa comparando los resultados de la prueba con los de un patrón de referencia (*gold-standard*), que identifica el diagnóstico verdadero. Para pruebas con resultados dicotómicos (ej. presencia-ausencia de enfermedad) la evaluación se concreta en distintos indicadores de validez: sensibilidad, especificidad y valores predictivos positivo y negativo. La sensibilidad y la especificidad son características intrínsecas de la prueba diagnóstica, mientras que los valores predictivos dependen también de la prevalencia o probabilidad preprueba de la enfermedad a estudio. Los cocientes de probabilidades son índices resumen de la sensibilidad y la especificidad, independientes de la probabilidad preprueba, que pueden usarse en la predicción de la probabilidad postprueba. Estos mismos estimadores pueden ser aplicados a pruebas con resultados discretos con más de dos categorías o continuos, si se establece un punto de corte o umbral diagnóstico. Las curvas ROC permiten explorar la capacidad diagnóstica de la prueba en sus distintos valores, de manera que podamos conocer su validez global y seleccionar el punto o puntos de corte más adecuados.

Es preciso tener en cuenta que la información que disponemos sobre la validez de las pruebas diagnósticas procede de estudios realizados en muestras de población. Por lo tanto, las estimaciones obtenidas en dichos estudios están sujetas a variabilidad aleatoria y, si los estudios han sido diseñados incorrectamente, a sesgos.

La fiabilidad de una prueba viene determinada por la estabilidad de sus mediciones cuando se repite en condiciones similares. La variabilidad de las mediciones va a estar influida por múltiples factores que interesa conocer y controlar. Entre ellos, tiene especial importancia distinguir las variaciones de interpretación intraobservador e interobservador. La fiabilidad puede ser evaluada para resultados discretos nominales mediante el índice *kappa*, para resultados discretos ordinales mediante el índice *kappa* ponderado y para resultados continuos mediante el coeficiente de correlación intraclase y el método de Bland-Altman.

Diagnóstico médico y pruebas diagnósticas

El diagnóstico es un proceso dinámico que se inicia con la anamnesis, en el que el médico comienza a emitir hipótesis sobre lo que le pasa al enfermo, hipótesis que son contrastadas y aceptadas o rechazadas provisionalmente. Esta misma dinámica se repite a lo largo de la exploración física, cuando se analizan los resultados de las pruebas complementarias e, incluso, cuando ya se ha instaurado el tratamiento[(1)].

El proceso diagnóstico se sustenta sobre un modelo probabilístico, en el que cada uno de sus pasos se traduce en una modificación del grado de certeza con el que emitimos el diagnóstico. Este dependerá, pues, no sólo del nivel de conocimientos clínicos y epidemiológicos del médico, sino también de su capacidad para concretarlos en un simple cálculo de probabilidades[(2,3)].

Veamos un ejemplo de aproximación probabilística al diagnóstico. Una causa común de fiebre en el lactante es la infección del tracto urinario. Por la información disponible en la literatura, el médico que atiende a un lactante con fiebre en un servicio de urgencias puede realizar una estimación previa de la probabilidad de que tenga una infección urinaria del 5,3%[(4)]. A partir de los datos obtenidos de la anamnesis y de la exploración, el médico puede incrementar su certidumbre. Así, si se trata de un lactante de sexo femenino con fiebre igual o superior a 39º C, la probabilidad ascendería a un 16,9%[(4)]. Si el médico recurre a la búsqueda de leucocituria en orina, la detección de 10 o más leucocitos por mm^3, le permitiría estimar una probabili-

[1]Servicio de Pediatría. Unidad de Investigación. Hospital Virgen de la Concha. Zamora. [2]Clinamat-Medycsa. Madrid.
Correspondencia: Carlos Ochoa Sangrador. Unidad de Investigación. Hospital Virgen de la Concha. Avd. Requejo 35. 49022 Zamora.

dad de infección del 56,4%[5]. Estas estimaciones de probabilidad van a orientar al médico en su actitud terapéutica, pero ninguna de ellas le facilita una seguridad absoluta sobre la existencia de infección urinaria, que sólo podrá ser confirmada por el hallazgo de un cultivo de orina positivo. En este ejemplo, el urocultivo es considerado un patrón de referencia o "patrón oro" (*gold standard*), a él se le asigna una validez absoluta (estimación de probabilidad del 100%), por lo que su resultado nos permite distinguir a los pacientes con infección urinaria de los que no la tienen. Si no disponemos del resultado del patrón de referencia nuestra mejor aproximación al diagnóstico nos la ofrecen la leucocituria y los otros datos clínicos. La contribución de cada uno de ellos al diagnóstico depende del grado de acuerdo que hayan mostrado previamente con el patrón de referencia. De este grado de acuerdo obtenemos las estimaciones concretas de probabilidad.

Desde un punto de vista funcional, podemos considerar prueba diagnóstica a cualquier procedimiento realizado para confirmar o descartar un diagnóstico o incrementar o disminuir su verosimilitud. En nuestro ejemplo anterior, tanto la leucocituria, como los otros datos clínicos se comportan como pruebas diagnósticas, ya que permiten modificar la probabilidad de un determinado diagnóstico.

El principio fundamental de las pruebas diagnósticas reside en la creencia de que los individuos que tienen una enfermedad son distintos de los que no la tienen, y que las pruebas diagnósticas permiten distinguir a los dos grupos. Las pruebas diagnósticas, para ser perfectas, requerirían que 1) todos los individuos sin la enfermedad tuvieran un valor uniforme en la prueba (habitualmente normal), 2) que todos los individuos con la enfermedad tuvieran un valor uniforme pero distinto en la prueba (habitualmente anormal) y 3) que no hubiera resultados indeterminados imposibles de asignar al mostrado por los enfermos o por los sanos. Pero en la práctica, resulta excepcional que estos requisitos se cumplan a la perfección. Existen variaciones en los resultados de las pruebas debidas a insuficiente fiabilidad de las mismas o a la existencia de heterogeneidad en las características de la población enferma y sana, que condicionan su validez. No obstante, conocer las características y las limitaciones de las pruebas diagnósticas le permite al médico tomar decisiones cuantificando el grado de certeza existente en sus juicios diagnósticos.

La calidad de una prueba diagnóstica depende, en primer lugar, de su capacidad para producir los mismos resultados cada vez que se aplica en similares condiciones, y en segundo lugar, de que sus mediciones reflejen exactamente el fenómeno que se intenta medir. Dicho de otro modo, una prueba diagnóstica debe ser fiable y válida. 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 éstas no sean válidas, tampoco tiene ninguna utilidad. Además de su fiabilidad y validez, la utilidad de una prueba también depende de su rendimiento clínico y de su coste. Una prueba muy válida y fiable, pero cuya contribución al diagnóstico apenas modifique la actitud del médico o cuya ejecución tenga un coste excesivo tendrá una escasa utilidad. El análisis de estos aspectos se abordarán en otro capítulo de esta serie.

Validez de las pruebas diagnósticas

Evaluación de la validez de las pruebas diagnósticas. Patrón de referencia

Una prueba diagnóstica será válida si es capaz de medir correctamente el fenómeno que pretende estudiar. Pero para poder evaluar la validez de una prueba diagnóstica se requiere un patrón de referencia o "patrón oro" (*gold standard*) que refleje fielmente la característica a medir. Para el escenario diagnóstico más simple, en el que el fenómeno a medir es la presencia o ausencia de una enfermedad, el patrón de referencia tendrá que clasificar perfectamente a la población enferma y a la sana. Cuanto mayor grado de acuerdo tenga la prueba diagnóstica con la prueba o patrón de referencia más válida será.

Aunque asumamos que el patrón de referencia tiene una validez absoluta, es frecuente que esta validez no sea perfecta o que no haya podido ser evaluada. A menudo, se asigna dicho papel a la prueba diagnóstica disponible con la que existe mayor experiencia o a la que ha demostrado, hasta un momento dado, mayor validez. La mayoría de los estudios de evaluación de pruebas diagnósticas tratan de comparar esa prueba estándar con una nueva prueba que presenta ventajas en cuanto a rendimiento[6], sencillez, rapidez, coste, seguridad, etc.[6-10].

En ocasiones no se dispone de ninguna prueba de referencia, por la naturaleza del concepto a medir o por la ausencia de conocimiento suficiente. En estas situaciones resulta útil recurrir a criterios diagnósticos diseñados por expertos[11,12] o resultantes de un conjunto de pruebas agrupadas. Cuando resulte imposible establecer un patrón de referencia, se tratará al menos de valorar la fiabilidad de la prueba estudiada y su concordancia con otras pruebas alternativas.

Es importante tener en cuenta que un "patrón oro" puede ser imperfecto y en ese caso sus defectos van a influir en la evaluación de la prueba diagnóstica. Aunque podemos examinar la fiabilidad del "patrón oro", su validez debe ser asumida por cuestiones de operatividad, al menos provisionalmente, hasta que sea sustituido por otra prueba, a la luz del avance del conocimiento o la tecnología. No obstante, si se tiene información sobre los sesgos que provoca un patrón de referencia, éstos pueden ser corregidos mediante modelos matemáticos[13].

Sensibilidad y especificidad

Consideremos el escenario diagnóstico más simple, en el que tanto el patrón de referencia como la prueba diagnóstica clasifican a los pacientes en dos grupos, en función de la presencia o ausencia de un síntoma, signo o enfermedad. Veamos un ejemplo. En el diagnóstico de infección urinaria es frecuente que se

Tabla I Tabla de contingencia 2 x 2 para la evaluación de una prueba diagnóstica

		Patrón de referencia +	Patrón de referencia –	
Prueba diagnóstica	+	Verdaderos positivos (a)	Falsos positivos (b)	a+b
	–	Falsos negativos (c)	Verdaderos negativos (d)	c+d
		a+c	b+d	Total=a+b+c+d

Claves:

a Verdaderos positivos (VP): enfermos con la prueba positiva
b Falsos positivos (FP): no enfermos con la prueba positiva
c Falsos negativos (FN): enfermos con la prueba negativa
d Verdaderos negativos (VN): no enfermos con la prueba negativa
a+c Casos con patrón de referencia positivo (enfermos)
b+d Casos con patrón de referencia negativo (no enfermos)
a+b Casos con la prueba diagnóstica positiva
c+d Casos con la prueba diagnóstica negativa

recurra a la detección de leucocitos en orina como prueba diagnóstica de infección urinaria, ya que la presencia de leucocitos en orina incrementa la verosimilitud de que el paciente tenga una infección urinaria[14]. En este caso el urocultivo, cuya positividad confirma la existencia de infección (para orinas obtenidas por técnica estéril), es el patrón de referencia[14]. Tanto la prueba diagnóstica (leucocituria) como el patrón de referencia (urocultivo) tratan de clasificar a los pacientes en dos grupos: enfermos o sanos.

En este escenario, el grado de acuerdo entre la prueba diagnóstica y el patrón de referencia puede ser representado en una tabla de contingencia (Tabla I). Habitualmente se sitúa en las columnas el resultado de la prueba de referencia que clasifica a los sujetos en enfermos y sanos (patrón de referencia positivo o negativo) y en las filas el resultado de la prueba diagnóstica (prueba positiva o negativa). En las cuatro casillas de la tabla se recogen el recuento de casos que cumplen las características señaladas en las cabeceras de sus columnas y filas correspondientes. De izquierda a derecha y de arriba abajo estas casillas (a,b,c,d) contienen los verdaderos positivos (VP), casos con patrón de referencia y prueba diagnóstica positivos, los falsos positivos (FP), casos con patrón de referencia negativo y prueba diagnóstica positiva, los falsos negativos (FN), casos con patrón de referencia positivo y prueba diagnóstica negativa y los verdaderos negativos (VN), casos con patrón de referencia y prueba diagnóstica negativos.

Se puede hacer una aproximación a la validez de la prueba diagnóstica calculando la proporción de aciertos, esto es, la proporción de pacientes con patrón de referencia positivo o negativo (enfermos o sanos) que son correctamente diagnosticados por la prueba. Del total de observaciones (a+b+c+d) son aciertos los verdaderos positivos (a) y los verdaderos negativos (d), de manera que la validez global de la prueba corresponde a:

$$\text{Validez} = \frac{VP + VN}{\text{Total}} = \frac{a + d}{a + b + c + d}$$

Pero conocer la proporción global de aciertos tiene menos interés que conocer la proporción de aciertos entre la poblaciones enferma y sana. Ambas proporciones se pueden calcular en la tabla de contingencia comprobando los aciertos por columnas (en dirección vertical). Estas proporciones definen las características operacionales de las pruebas diagnósticas: sensibilidad y especificidad.

La **sensibilidad** (Se) es la probabilidad de que la prueba dé positiva si la condición de estudio está presente (paciente enfermo o con patrón de referencia positivo). También se puede definir como la proporción de verdaderos positivos respecto al total de enfermos.

$$\text{Sensibilidad:} \quad Se = \frac{VP}{\text{Enfermos}} = \frac{a}{a + c}$$

La **especificidad** (Es) es la probabilidad de que la prueba dé negativa si la enfermedad está ausente (paciente sano o con patrón de referencia negativo). También se puede definir como la proporción de verdaderos negativos respecto al total de sujetos sanos.

$$\text{Especificidad:} \quad Es = \frac{VN}{\text{Sanos}} = \frac{d}{b + d}$$

Veamos un ejemplo de la aplicación de estos indicadores sobre los datos de un estudio en el que se evaluaron distintos elementos del análisis de orina para el diagnóstico de infección urinaria en pacientes pediátricos ambulatorios[14]. En la tabla II se presentan los resultados del test de la estearasa leucocitaria mediante tira reactiva (prueba diagnóstica) con respecto a los del urocultivo (patrón de referencia). La prueba diagnóstica será válida si es capaz de discriminar entre enfermos (urocultivo positivo) y no enfermos (urocultivo negativo). Las características operativas en nuestro ejemplo serán:

$$Se = \frac{VP}{\text{Enfermos}} = \frac{a}{a + c} = \frac{81}{102} = 0{,}79$$

$$Es = \frac{VN}{\text{Sanos}} = \frac{d}{b + d} = \frac{427}{587} = 0{,}72$$

Basándonos en los resultados de este estudio podríamos estimar que el 79% de los pacientes con urocultivo positivo tendrán un test de la estearasa leucocitaria positivo, mientras que el 72% de los pacientes con urocultivo negativo tendrán un test negativo.

Sensibilidad o especificidad

Sensibilidad y especificidad son propiedades intrínsecas de las pruebas diagnósticas, ya que no dependen de la probabilidad preprueba o prevalencia del fenómeno que se estudia. Tienen una utilidad preprueba, informan de la validez de la prueba an-

Tabla II Tabla de contingencia de la evaluación del test de la estearasa leucocitaria para el diagnóstico de la infección urinaria (urocultivo positivo)[14]. Ver la interpretación de las casillas en la tabla I

		Urocultivo +	Urocultivo –	
Test dela estearasa	+	81	160	241
leucocitaria	–	21	427	448
		102	587	689

tes de realizarla. La sensibilidad considera la validez de la prueba entre los enfermos y la especificidad la validez de la prueba entre los sanos. Aunque ambas características son importantes, en determinadas ocasiones se prefieren pruebas más sensibles y en otras pruebas más específicas. Cuando empleamos una prueba diagnóstica como técnica de cribado poblacional nos interesa una prueba muy sensible, para que no se nos escape ningún enfermo (falso negativo). Cuando se trata de confirmar un diagnóstico nos interesará una prueba con alta especificidad, para tratar de reducir el riesgo de catalogar como enfermo a un sujeto sano (falso positivo). Como se verá al hablar del umbral diagnóstico, la sensibilidad y la especificidad de las pruebas diagnósticas con resultados discretos con más de dos categorías o continuos, tienen una relación inversa según donde se sitúe el punto de corte diagnóstico.

Intervalos de confianza

Los valores de sensibilidad y especificidad son estimaciones puntuales obtenidas en estudios realizados con muestras de población, por lo que están sujetas a variabilidad aleatoria. Para expresar correctamente estas estimaciones deben calcularse sus **intervalos de confianza**, por métodos exactos o, si las muestras son suficientemente grandes, por aproximación a la normal[15]. Las fórmulas de los intervalos de confianza al 95% (IC 95%) por aproximación a la normal y los cálculos para el ejemplo del test de la estearasa leucocitaria[14] son los siguientes:

$$\text{IC 95\% de la sensibilidad: } Se \pm 1{,}96 \sqrt{\frac{Se \times (1 - Se)}{\text{Enfermos}}}$$

$$\text{IC 95\% de la especificidad: } Es \pm 1{,}96 \sqrt{\frac{Es \times (1 - Es)}{\text{Sanos}}}$$

En nuestro ejemplo:

$$\text{Se: } 0{,}79 \pm 1{,}96 \sqrt{\frac{0{,}79 \times (1 - 0{,}79)}{102}} \Rightarrow [0{,}71 - 0{,}86]$$

$$\text{Es: } 0{,}71 \pm 1{,}96 \sqrt{\frac{0{,}71 \times (1 - 0{,}71)}{}} \Rightarrow [0{,}67 - 0{,}74]$$

Teniendo en cuenta estos intervalos de confianza, estimaremos que de los pacientes con urocultivo positivo entre un 71 y un 86% tendrán el test de la estearasa leucocitaria positivo, mientras que de los pacientes con urocultivo negativo tendrán el test de la estearasa leucocitaria negativo entre un 67 y un 74%. Para poder comparar las características operativas de distintas pruebas diagnósticas no basta con las estimaciones puntuales, sino que tendremos que considerar sus intervalos de confianza.

Kappa ponderado de Se y Es. Medición calibrada de la calidad de la prueba

Como hemos dicho anteriormente sensibilidad y especificidad expresan porcentajes de acuerdo entre la prueba diagnóstica y el patrón de referencia. Cuando las calculamos asumimos que todo el acuerdo encontrado se debe a la bondad de la prueba, sin embargo, parte del acuerdo puede ser debido al azar. Aunque la prueba diagnóstica no tuviera nada que ver con el fenómeno de estudio, por azar acertaría en algunas observaciones. Por lo tanto, sensibilidad y especificidad son medidas de acuerdo no calibradas[16].

Para conocer el verdadero grado de acuerdo debido a la bondad de la prueba debe descontarse el debido al azar. Este puede ser estimado en la tabla de contingencia para cada casilla (Tabla I), considerando la ley multiplicativa de la probabilidad para sucesos independientes[17]. Si la prueba diagnóstica y el patrón de referencia son sucesos independientes, la probabilidad de que una observación sea por azar un verdadero positivo (casilla a) es igual a la probabilidad de que la prueba sea positiva multiplicado por la probabilidad de que el patrón de referencia también lo sea. Si multiplicamos la probabilidad resultante por el número total de observaciones obtenemos el recuento de verdaderos positivos (casilla a) esperado por azar. Este cálculo se puede simplificar como el producto de los marginales de la fila y columna correspondientes dividido por el total:

$$a_{esperados} = \left[\frac{a + b}{\text{Total}} \times \frac{a + c}{\text{Total}}\right] \times \text{Total} = \frac{(a + b) \times (a + c)}{\text{Total}}$$

Puede verse como la sensibilidad esperada por azar es igual a la probabilidad de tener una prueba diagnóstica positiva:

$$Se_{esperada} = \frac{\frac{(a + b) \times (a + c)}{\text{Total}}}{a + c} = \frac{a + b}{\text{Total}}$$

Los coeficientes *kappa* ponderados de sensibilidad (κ_{Se}) y especificidad (κ_{Es}) son medidas calibradas del grado de acuerdo, que descuentan las partes de sensibilidad y especificidad debidas al azar. Sus fórmulas y cálculos para el ejemplo de la estearasa leucocitaria[14] son los siguientes:

$$\text{*kappa* ponderado de sensibilidad } \kappa_{Se} = \frac{Se - \frac{\cancel{a + b}}{\text{Total}}}{1 - \frac{\cancel{a + b}}{}}$$

kappa ponderado de especificidad $\kappa_{Es} = \dfrac{Es - \frac{c+d}{Total}}{1 - \frac{c+d}{Total}}$

En nuestro ejemplo:

$$\kappa_{Se} = \frac{0,79 - 0,34}{1 - 0,34} = 0,68$$

$$\kappa_{Es} = \frac{0,72 - 0,65}{1 - 0,65} = 0,20$$

A la hora de documentar la validez de una prueba diagnóstica, junto a las estimaciones de sensibilidad y especificidad deben proporcionarse también estos coeficientes. Podemos observar cómo en nuestro ejemplo estos coeficientes presentan diferencias con la sensibilidad y especificidad no calibradas (0,79 y 0,72 respectivamente). En concreto, los bajos valores encontrados para el κ_{Es} (0,20) cuestionan la especificidad de la prueba[18,19].

Valores predictivos

Hasta ahora hemos hablado de la validez de las pruebas diagnósticas en cuanto a su concordancia con el patrón de referencia. Pero habitualmente cuando realizamos una prueba diagnóstica desconocemos el resultado del patrón de referencia, por ello, una vez conocido el resultado de la prueba, lo que nos interesa es estimar la probabilidad de que su diagnóstico sea correcto. Ni la sensibilidad ni la especificidad nos aportan esa información. Para ello debemos calcular los valores predictivos.

El **valor predicitivo positivo** (VPP) es la probabilidad de tener la condición de estudio (enfermedad o patrón de referencia positivo) si la prueba ha sido positiva. También puede ser definido como la proporción de verdaderos positivos respecto al total de pruebas positivas.

El **valor predictivo negativo** (VPN) es la probabilidad de no tener la condición de estudio (enfermedad ausente o patrón de referencia negativo) si la prueba ha sido negativa. También puede ser definido como la proporción de verdaderos negativos respecto al total de pruebas negativas.

Los valores predictivos pueden calcularse a partir de la tabla de contingencia (Tabla I) comprobando los aciertos por filas (en dirección horizontal). Las fórmulas de cálculo y su aplicación en el ejemplo de la estearasa leucocitaria[14] son las siguientes:

$$\text{Valor predictivo positivo VPP} = \frac{VP}{\text{Pruebas positivas}} = \frac{a}{a+b}$$

$$\text{Valor predictivo negativo VPN} = \frac{VP}{\text{Pruebas negativas}} = \frac{d}{c+d}$$

En nuestro ejemplo:

$$VPP = \frac{81}{241} = 0,33$$

$$VPN = \frac{427}{448} = 0,95$$

Basándonos en los resultados del estudio podríamos estimar que el 33% de los pacientes con el test de la estearasa leucocitaria positivo tendrán un urocultivo positivo, mientras que el 95% de los pacientes con el test negativo tendrán el urocultivo negativo.

Los valores predictivos tienen utilidad postprueba, ya que informan de la probabilidad de enfermedad una vez realizada la prueba y conocido su resultado (probabilidad postprueba). Sin embargo, sus predicciones tienen una validez limitada, porque dependen de la prevalencia del fenómeno en estudio en la población donde se aplica (probabilidad preprueba). Si la probabilidad preprueba en el entorno donde la vamos aplicar es diferente de la existente en el estudio que evaluó, la utilidad de la prueba, los valores predictivos no son válidos y deben ser ajustados.

Para ajustar los valores predictivos a cualquier prevalencia, se puede recurrir a formulaciones matemáticas basadas en el teorema de Bayes, a partir de los valores de sensibilidad, especificidad y de la probabilidad preprueba (Ppre):

$$VPP = \frac{Se \times Ppre}{Se \times Ppre + (1 - Es) \times (1 - Ppre)}$$

$$VPN = \frac{Es \times (1 - Ppre)}{(1 - Se) \times Ppre + Es \times (1 - Ppre)}$$

donde Ppre es la probabilidad preprueba o prevalencia del fenómeno en estudio, que se puede estimar en la tabla de contingencia (Tabla I):

$$Ppre = \frac{a+b}{Total}$$

Si quisiéramos aplicar el test de la estearasa leucocitaria para el cribado de infección urinaria en una población de lactantes sanos, tendríamos que ajustar los valores predictivos calculados en el estudio previo[14], porque la prevalencia esperada de infección urinaria en lactantes sanos será previsiblemente muy inferior a la encontrada en aquel estudio. Si consideramos la información disponible en la literatura, podemos estimar que la probabilidad de encontrar un urocultivo positivo (bacteriuria asintomática) en lactantes sanos es aproximadamente del 1%[20-22]. Ajustando a dicha prevalencia los valores predictivos quedarían:

Ppre=0,01

$$VPP = \frac{0,79 \times 0,01}{0,79x0,01+(1-0,72)x(1-0,01)} = \frac{0,0079}{0,0079+0,2772} = 0,027$$

$$VPN = \frac{0,72 \times (1-0,01)}{(1-0,79)x0,01+0,72x(1-0,01)} = \frac{0,7128}{0,0021+0,7128} = 0,997$$

Utilizando estos valores podríamos estimar que el 2,7% de los lactantes sanos con el test de la estearasa leucocitaria positivo tendrán un urocultivo positivo, mientras que el 99,7% de los pacientes con el test negativo tendrán el urocultivo negativo. Como puede comprobarse en este ejemplo, cuando la prevalencia de la enfermedad es muy baja, el valor predictivo positivo es bajo. Este hecho ocurre incluso con pruebas diagnósticas altamente sensibles y específicas. Por ello, al aplicar pruebas diagnósticas para cribado de grupos de población general es frecuente que muchos de los sujetos con pruebas positivas sean falsos positivos.

La prevalencia puede interpretarse como la probabilidad esperada de tener el fenómeno en estudio (enfermedad) antes de realizar la prueba diagnóstica. Esta probabilidad preprueba se comporta como un punto de partida en la predicción diagnóstica. Los valores predictivos son estimaciones revisadas de la probabilidad previa, distintas según sea el resultado de la prueba positivo o negativo, por lo que las denominamos probabilidades postprueba. La diferencia que exista entre las probabilidades preprueba y postprueba informa de la utilidad que tiene una determinada prueba diagnóstica. A mayor diferencia entre una y otra probabilidad mayor contribución de la prueba al proceso diagnóstico.

Cocientes de probabilidades

Una forma alternativa de describir el comportamiento de una prueba diagnóstica en este proceso son los cocientes de probabilidades. El **cociente de probabilidades** (CP) para un determinado resultado de una prueba diagnóstica está definido como la probabilidad de dicho resultado en presencia de enfermedad dividida por la probabilidad de dicho resultado en ausencia de enfermedad. Los CP resumen información de la sensibilidad y de la especificidad e indican la capacidad de la prueba para incrementar o disminuir la verosimilitud de un determinado diagnóstico. A partir de los CP se pueden calcular las probabilidades postprueba (valores predictivos) para cualquier prevalencia.

El CP del resultado positivo de una prueba diagnóstica, CP a favor o positivo (CP+), indica cuánto más probable es que la prueba sea positiva en un paciente enfermo respecto a uno sano. Este CP+ se puede calcular a partir de la sensibilidad y la especificidad con la fórmula:

$$CP+ = \frac{Se}{1 - Es}$$

La fórmula del CP del resultado negativo de una prueba, CP en contra o negativo (CP–) es:

$$CP- = \frac{1- Se}{Es}$$

En el ejemplo de la estearasa leucocitaria (Tabla II)[14] los CP+ y CP– serán:

$$CP+ = \frac{0,79}{1 - 0,72} = 2,82$$

$$CP- = \frac{1 - 0,79}{0,72} = 0,29$$

Los CP adoptan valores entre 0 e infinito, siendo el valor nulo el 1 (no modifica las *odds* previas). Cuanto más elevado sea el CP por encima de 1 más se incrementará la probabilidad del diagnóstico, cuanto más bajo sea el CP por debajo de 1 más disminuirá la probabilidad del diagnóstico. Al igual que las otras características operativas, los CP son estimadores obtenidos en muestras de población, por lo que deben calcularse con sus intervalos de confianza[23].

La principal utilidad de los CP es que permiten calcular la probabilidad postprueba a partir de cualquier prevalencia. Para poder operar con los CP en el cálculo de probabilidades, éstas deben transformarse en ventajas (*odds*). Las ventajas u *odds* se calculan dividiendo las probabilidades por sus complementarios (P/1-P). Los pasos a seguir en el cálculo de la probabilidad postprueba son: 1) transformar la probabilidad preprueba en *odds* preprueba, 2) multiplicar la *odds* preprueba por el CP del resultado encontrado en la prueba (CP+ o CP–) con lo que se obtiene la odds postprueba, 3) transformar la *odds* postprueba en probabilidad.

Veamos estos pasos en el cálculo de la probabilidad postprueba (Ppost) de tener urocultivo positivo, tras obtener un resultado positivo al aplicar el test de la estearasa leucocitaria en la orina de un lactante sano (prevalencia o probabilidad preprueba estimada = 0,01):

$$\text{Odds preprueba} = \frac{Ppre}{1 - Ppre} = \frac{0,01}{1 - 0,01} = 0,01$$

Odds postprueba = CP+ x Odds preprueba= 2,82x0,01=0,028

$$Ppost = \frac{\text{Odds postprueba}}{1 + \text{Odds postprueba}} = \frac{0,028}{1 + 0,028} = 0,027$$

En conclusión, solo el 2,7% de los lactantes que tengan una tira reactiva positiva tendrán infección urinaria. Esta cifra es la misma que la calculada anteriormente con la fórmula bayesiana del valor predictivo positivo.

Una de las ventajas de los CP es que si la prueba tiene más de 2 resultados posibles, se puede calcular un CP para cada uno de ellos, permitiéndonos interpretar la contribución al diagnóstico de cada resultado. Otra de las ventajas radica en que los CP facilitan el cálculo de las modificaciones de probabilidad obtenidas al aplicar en serie varias pruebas diagnósticas, recurso frecuentemente empleado en la práctica clínica y en los estudios de análisis de decisión.

Tabla III Distribución de 885 pacientes entre 3 y 36 meses de edad, con fiebre elevada (≥ 39 °C), atendidos en servicios de urgencias hospitalarios, en función del recuento de leucocitos en sangre y de la presencia o ausencia de bacteriemia detectada en hemocultivo[24]

Recuento de leucocitos en sangre nº/mm³	*Sin bacteriemia*	*Con bacteriemia*
5.000-9.999	366	2
10.000-14.999	293	7
15.000-19.999	130	7
20.000-24.999	44	4
≥ 25.000	26	6
Total (885)	859	26

Umbral diagnóstico (punto de corte). Curvas ROC

Hasta el momento hemos considerado un escenario diagnóstico simple en el que la prueba diagnóstica solo tenía dos resultados posibles: positivo o negativo. Sin embargo este escenario no se ajusta a las características de las pruebas diagnósticas cuyos resultados se miden en escalas discretas ordinales o continuas. En estas circunstancias la solución aparentemente más sencilla es establecer un umbral diagnóstico o punto de corte entre todos los valores posibles que nos permita discriminar entre los resultados positivos y negativos. Pero la elección de dicho umbral diagnóstico no resulta fácil, ya que tiene importantes implicaciones sobre la utilidad diagnóstica de la prueba.

Lo habitual en la práctica clínica es constatar que existe cierto grado de solapamiento entre los resultados de las pruebas diagnósticas de la población enferma y sana. Rara vez contamos con un punto de corte que discrimine totalmente a ambos grupos. Por ello, las características operacionales de las pruebas diagnósticas, sensibilidad y especificidad, van a cambiar según donde pongamos el punto de corte. Habitualmente, ambas características tendrán una relación inversa. Si tratamos de incrementar la sensibilidad, llevando el punto de corte hacia valores normales, disminuiremos la especificidad, y en sentido contrario si tratamos de aumentar la especificidad, llevando el punto de corte hacia valores anormales, reduciremos la sensibilidad.

Veamos un ejemplo para ilustrar esta cuestión. Consideremos el recuento total de leucocitos en sangre periférica como una prueba diagnóstica de la presencia de bacteriemia en niños con fiebre elevada. El riesgo de bacteriemia será mayor en los pacientes con mayores recuentos, pero también existirán algunos casos entre los pacientes con recuentos bajos. En la tabla III podemos observar las bacteriemias detectadas según el recuento de leucocitos en un estudio llevado a cabo en 885 niños de 3 a 36 meses de edad, con fiebre elevada (≥ 39 ºC), atendidos en servicios de urgencias hospitalarios[24]. Es evidente la existencia de solapamiento en los valores que presentan los pacientes con y sin bacteriemia. En la tabla IV se presentan las variaciones en las características operativas de la prueba del ejemplo anterior[24] empleando 5 puntos de corte distintos. Puede observarse como la sensibilidad (porcentaje de verdaderos positivos) va disminuyendo y la especificidad (porcentaje de verdaderos negativos) va aumentando, a medida que elevamos el punto de corte.

Pero antes de plantearnos la elección de un punto de corte, tenemos que determinar si los resultados de la prueba diagnóstica son capaces de discriminar entre la población que tiene y la que no tiene la característica de interés. Una forma de explorar la capacidad diagnóstica de las pruebas a lo largo de todos los posibles puntos de corte es el análisis de las características operativas de los receptores o curvas ROC (iniciales del término inglés original *Receiver Operating Characteristic*). Son una representación gráfica de la relación existente entre sensibilidad y especificidad para cada punto de corte posible.

Para confeccionar una curva ROC se deben calcular la sensibilidad y la especificidad para todos los posibles puntos de corte de la prueba diagnóstica. La curva se construye a partir de la representación de los distintos puntos de corte en una gráfica de dispersión, cuyos ejes de coordenadas vertical (y) y horizontal (x) corresponden a la sensibilidad y al complementario de la especificidad (proporción de falsos positivos). En la figura 1 podemos ver la curva ROC del recuento de leucocitos en sangre, construida con los datos de la tabla IV. Una prueba que discriminara perfectamente entre los dos grupos de pacientes, describiría una curva que coincidiría con los lados izquierdo y superior del gráfico. Una prueba que fuera totalmente inútil adoptaría la forma de una línea recta entre las esquinas inferior-izquierda y superior-derecha del gráfico (línea punteada de la figura 1). En la práctica, las curvas se situarán en una posición intermedia entre esas dos opciones.

El área bajo la curva representa la validez global de la prueba. Cuanto más se aproxima la curva a la esquina superior-izquierda del gráfico, mayor será esa área y mayor la validez de la prueba diagnóstica. Las curvas ROC nos permiten contrastar la capacidad diagnóstica de dos o más pruebas, comparando las áreas bajo las curvas de cada una de ellas[25,26]. Al igual que otras características de las pruebas diagnósticas, las curvas ROC son estimaciones poblacionales obtenidas a partir de muestras, por lo que están sujetas a error aleatorio. Este error puede ser representado construyendo bandas de confianza para las curvas[27].

Una vez establecida la validez global de la prueba diagnóstica, podemos plantearnos la elección del mejor punto de corte para su uso clínico. El mejor punto de corte no es aquel en el que se producen menos errores de clasificación, ya que también hay que tener en cuenta otros aspectos que dependen de las condiciones existentes en el entorno donde va a ser aplicada la prueba. Estos aspectos son fundamentalmente el coste (no solamente económico) de los resultados falsos, tanto positivos como negativos, los beneficios de las clasificaciones correctas y la prevalencia esperada de la condición de estudio. La valoración de todos estos aspectos resulta compleja y no está exenta de un cierto grado de subjetividad, aunque se hayan desarrollado distintas

Tabla IV Variaciones en las características operativas del recuento de leucocitos en sangre para el diagnóstico de bacteriemia en niños de 3 a 36 meses con fiebre elevada, considerando 5 puntos de corte con los datos de la tabla III[24]. Se presentan los verdaderos positivos (VP), los falsos positivos (FP), los verdaderos negativos (VN), los falsos negativos (FN), la sensibilidad (Se), la especificidad (Es) y el complementario de la especificidad (1-Es)

Recuento de leucocitos / mm³	*VP*	*FP*	*VN*	*FN*	*Se*	*Es*	*1-Es*
≥ 5.000	26	859	0	0	1,00	0,00	1,00
≥ 10.000	24	493	366	2	0,92	0,43	0,57
≥ 15.000	17	200	659	9	0,65	0,77	0,23
≥ 20.000	10	70	789	16	0,38	0,92	0,08
≥ 25.000	6	26	833	20	0,23	0,97	0,03

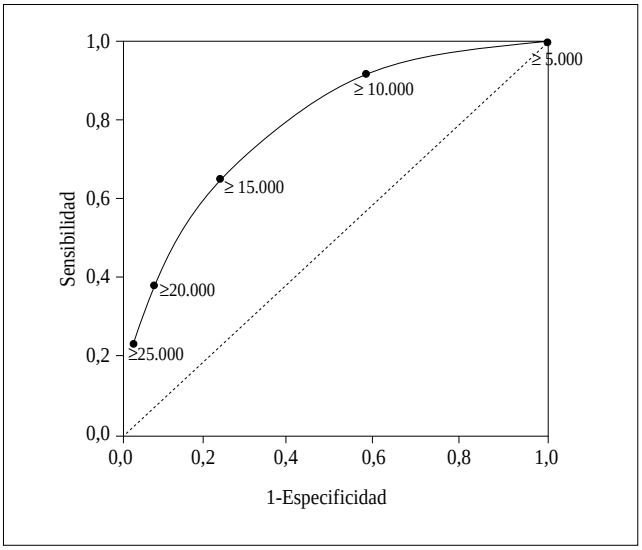

Figura 1. Características operativas (curva ROC) del recuento de leucocitos en sangre como prueba diagnóstica de bacteriemia. Construida con los datos de la tabla IV[24].

herramientas para su sistematización[26,28,29]. En general, nos interesará escoger un punto de corte con alta sensibilidad cuando la prueba vaya a ser aplicada como técnica de cribado poblacional, para que no se nos escape ningún enfermo (falso negativo), siempre que el coste de la identificación de falsos positivos no sea elevado y pueda ser reducido aplicando, en un segundo paso, otras pruebas más específicas.

Errores en el diseño de los estudios de evaluación de pruebas diagnósticas

Cuando un clínico quiere conocer la validez de una determinada prueba diagnóstica consulta los estudios publicados en que ha sido valorada, asumiendo las estimaciones de las características operativas de la prueba obtenidas en los mismos. Sin embargo, si dichos estudios están mal diseñados, realizados o presentados podemos sacar conclusiones erróneas sobre su validez y utilidad[30-32].

A la hora de analizar la evidencia que un determinado estudio aporta sobre la validez de una prueba diagnóstica, debemos plantearnos las siguiente cuestiones[32-37]:

¿Ha sido comparada la prueba con un verdadero patrón de referencia (gold standard)?

Tal y como comentamos en el apartado de evaluación de las pruebas diagnósticas, el patrón de referencia empleado tiene que contar con una validez contrastada o, al menos, aceptada por consenso. La utilización de un patrón de referencia defectuoso puede introducir sesgos en las estimaciones de validez de la prueba diagnóstica.

En relación con el patrón de referencia, resulta también importante considerar si es capaz de clasificar el estado de enfermedad en todas las observaciones. En el caso de que existan observaciones con un diagnóstico indeterminado, si éstas son excluidas del análisis, se producirán estimaciones sesgadas de las características operativas de la pruebas diagnósticas. Este sesgo, conocido como *sesgo por exclusión de indeterminados*, ocasiona habitualmente sobrestimaciones de la sensibilidad y de la especificidad[38,39].

Otro sesgo, relacionado con el patrón de referencia, que debe tratar de evitarse es el *sesgo de incorporación*. Este sesgo ocurre cuando elementos de la prueba diagnóstica forman parte del patrón de referencia. Por ejemplo, si evaluamos el papel de la resonancia magnética en el diagnóstico de una enfermedad neurológica, y la resonancia magnética está incluida en la batería de pruebas que definen el diagnóstico de referencia de dicha enfermedad, tendremos un sesgo de incorporación. Hay que tener en cuenta que en esta situación la sensibilidad y la especificidad se sobrestiman.

¿La muestra estudiada incluye un apropiado espectro de sujetos?

Para poder asumir las características de la prueba obtenidas en un estudio, las características de la muestra deberían asemejarse a las de la población donde vamos a aplicarla. Aunque hemos considerado hasta el momento que la sensibilidad y la especificidad son características intrínsecas de las pruebas diagnósticas, que no dependen de la prevalencia de la enfermedad, en la práctica, podemos encontrarnos diferencias en función de

las características epidemiológicas de la muestra. Los resultados pueden ser distintos si la población sana tiene mayor o menor riesgo de tener la condición objeto de estudio o la población enferma está en diferente estadio evolutivo. Para controlar estos aspectos es preciso que los criterios de selección y las características clínicas y epidemiológicas de la muestra analizada estén claramente presentados. Asimismo, si se prevé la existencia de diferencias importantes, puede resultar necesario analizar el comportamiento de la prueba por subgrupos clínicos o demográficos, de manera que podamos escoger los estimadores que más se ajusten a nuestro entorno.

¿Se ha evitado el sesgo de secuencia o verificación diagnóstica?

El diseño del estudio debe tratar de garantizar que a todos los sujetos se les haya realizado, tanto la prueba diagnóstica, como el patrón de referencia. Los descriptores de validez de la prueba podrán ser calculados directamente a partir de los datos obtenidos cuando la prueba diagnóstica y la prueba de referencia son realizadas de forma simultánea (*diseño simultáneo*), siempre que en la muestra no se hayan excluido pacientes, en función del resultado de la prueba o de la existencia de mayor o menor riesgo de enfermedad. Pero esta estrategia simultánea, sin duda la más válida, resulta en ocasiones poco factible. Si la enfermedad es rara, la obtención de descriptores fiables requiere el estudio de grandes muestras y la realización de las pruebas diagnóstica y de referencia a muchos sujetos sanos. Por ello, se recurre con frecuencia a otros diseños.

Una opción más eficiente para la evaluación de pruebas diagnósticas es el *diseño retrospectivo*. En él, se determina en un primer paso la presencia o ausencia de enfermedad y en un segundo paso se realiza la prueba diagnóstica a dos submuestras representativas de los sujetos con y sin enfermedad. Con esta estrategia podemos calcular directamente la sensibilidad y la especificidad, pero los valores predictivos deben ser obtenidos con las fórmulas bayesianas, que trabajan con probabilidades condicionales[15].

Otra opción, que reproduce el proceder habitualmente utilizado en la práctica clínica, es el *diseño prospectivo*. En él, se aplica, en un primer paso, la prueba diagnóstica a una muestra de la población susceptible de estudio. A continuación se toman dos submuestras representativas de los sujetos con la prueba positiva y negativa, a las que se les realiza la prueba de referencia. Con esta estrategia, los valores predictivos se pueden calcular directamente, pero la sensibilidad y la especificidad deben ser estimadas por fórmulas bayesianas.

El principal *sesgo* en que se puede incurrir en estudios con un diseño prospectivo es el de *verificación diagnóstica*. Este ocurrirá cuando la probabilidad de que se les realice la prueba de referencia sea menor entre los sujetos con la prueba diagnóstica negativa y por lo tanto sea menos probable que éstos entren en el estudio. Este sesgo producirá sobrestimaciones de la sensibilidad e infraestimaciones de la especificidad.

¿Se ha evitado el sesgo de revisión?

Se puede incurrir en un sesgo de revisión cuando la interpretación de la prueba diagnóstica y del patrón de referencia no se hace de forma independiente. Si el resultado de una prueba es susceptible de interpretación subjetiva, puede verse influenciado por el conocimiento del diagnóstico o de características clínicas del paciente que lo hagan más o menos probable. Para poder garantizar la validez de las estimaciones, deben realizarse de forma ciega la prueba diagnóstica y el patrón de referencia.

¿Están los resultados del estudio convenientemente presentados?

Como los descriptores de validez de las pruebas diagnósticas son estimaciones puntuales de los verdaderos descriptores poblacionales, están sujetos a variabilidad aleatoria y, por lo tanto, deben proporcionarse con sus intervalos de confianza. Estos intervalos de confianza tendrán que ser considerados a la hora de interpretar la validez de una prueba diagnóstica. Si el tamaño muestral del estudio de evaluación es muy pequeño, los intervalos de confianza serán muy amplios, por lo que valoración de la validez de las pruebas diagnósticas puede quedar muy limitada.

Otro aspecto importante en la presentación de los resultados es la correcta utilización de las distintas herramientas disponibles para el análisis de la validez de las pruebas diagnósticas. En este sentido, interesa destacar la gran utilidad práctica que tienen herramientas como los cocientes de probabilidades y las curvas ROC.

¿Se ha considerado el papel de la prueba en el contexto del conjunto de posibles pruebas del proceso diagnóstico?

Es frecuente que durante el proceso diagnóstico recurramos a distintas pruebas diagnósticas, que pueden ser empleadas en paralelo o en serie con la prueba analizada en el estudio. En estas circunstancias, la validez de la prueba debe ser analizada en el contexto de todas las pruebas disponibles. La contribución de la prueba al diagnóstico, dependerá del momento en que se use, ya que el conocimiento del resultado de otras pruebas va a incidir en las probabilidades preprueba y postprueba. Para poder integrar convenientemente las distintas pruebas resultan útiles los CP, los árboles de decisión[40] y los modelos multivariantes[41].

¿Se ha evaluado la fiabilidad de la prueba diagnóstica?

Como se comentó en la introducción de este capítulo, la fiabilidad de una prueba diagnóstica es un requisito previo al de validez. Por lo tanto, en los estudios de evaluación de pruebas diagnósticas es conveniente incluir algún tipo de análisis de la consistencia de sus mediciones.

Fiabilidad de las pruebas diagnósticas

Hasta el momento hemos abordado el análisis de la validez de las pruebas diagnósticas. Pero la calidad de una prueba diag-

Tabla V Evaluación por parte de 2 médicos de las radiografías de tórax de 100 niños con sospecha de neumonía (datos figurados). Las casillas reflejan el recuento de casos en que hay acuerdo y desacuerdo

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

nóstica no depende exclusivamente de su validez, también depende de su fiabilidad.

La fiabilidad o consistencia 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.

Variables discretas nominales. Indice *kappa*

El índice kappa puede aplicarse a pruebas cuyos resultados sólo tengan dos categorías posibles o más de dos sin un orden jerárquico entre ellas. En la tabla V 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:

$$P_o = \frac{a + d}{Total} = \frac{4 + 84}{100} = 0,86\ (86\%)$$

Tabla VI Estimación de las observaciones esperadas por azar en la tabla de contingencia del ejemplo de la tabla V

		Médico A: Neumonía	Médico A: 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

Una concordancia del 86% podría ser interpretada como buena, sin embargo es preciso tener en cuenta que parte del acuerdo encontrado puede ser debido al azar. Tal y como se comentó al hablar de los índices *kappa* de sensibilidad y especificidad, 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 VI se presentan los cálculos para cada una de las casillas del ejemplo de la tabla V. Considerando estos recuentos estimados, la proporción de acuerdo esperada por azar sería:

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

Podemos constatar que existe acuerdo por azar en una elevada proporción de observaciones (79%). Si excluimos del análisis dichas observaciones, solo quedarán 7 observaciones concordantes (86-79=7) en un total de 21 observaciones (100-79=21), lo que supone un grado de acuerdo no debido al azar del 33% (7/21=0,33). 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 V) obtenemos:

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0,86 - 0,79}{1 - 0,79} = 0,33$$

lo que supone un grado de concordancia no debido al azar del 33%, 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

Tabla VII 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

de valores situados entre 0 y 1 se expone en la tabla VII[18,19]. Al igual que con otros estimadores poblacionales expuestos en este capítulo, los índices *kappa* se deben calcular con sus intervalos de confianza[19].

El índice kappa también puede ser aplicado a pruebas cuyos resultados tengan más de 2 categorías nominales, utilizando la misma metodología para el cálculo del acuerdo esperado por azar.

Variables discretas ordinales. Indice kappa ponderado.

El índice *kappa* ponderado debe emplearse cuando el resultado de la prueba analizada puede adoptar más de 2 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 VIII 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. Por ejemplo, si en nuestro ejemplo asignamos 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 sólo 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 distintas casillas se deben multiplicar por sus pesos respectivos. En la tabla IX podemos ver los pesos asignados en el ejemplo de la tabla VIII 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 los siguientes:

Tabla VIII Resultados de dos evaluaciones sucesivas, separadas por un corto período 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

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

$$P_o = \frac{1 \cdot (35 + 10 + 11) + 0{,}25 \cdot (8 + 9 + 12 + 15)}{100} = 0{,}64$$

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

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

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 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:

$$W_{i,j} = 1 - \frac{i - j}{k - 1}$$

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

Variables continuas

Desviación estándar intrasujetos

Cuando el resultado de una prueba se mide en una escala continua, podemos estimar el error de medición calculando la variabilidad existente entre medidas repetidas en los mismos sujetos. El parámetro que mejor refleja dicha variabilidad es la desviación estándar intrasujetos (excluyendo la observada entre su-

Tabla IX 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 VIII (ecuaciones de cada casilla)

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

Tabla X Resultados de dos mediciones repetidas de bilirrubina transcutánea (Jaundice-Meter 101, Minolta Air Shields), en la cara anterior del tórax en 20 recién nacidos ictéricos. Datos extraídos de un estudio más amplio[42]

Sujetos	*1ª medición*	*2ª medición*	*Diferencia*	*Media*
1	14	16	-2	15,0
2	14	14	0	14,0
3	17	17	0	17,0
4	14	15	-1	14,5
5	15	14	1	14,5
6	18	19	-1	18,5
7	16	16	0	16,0
8	12	12	0	12,0
9	19	19	0	19,0
10	9	10	-1	9,5
11	15	16	-1	15,5
12	18	18	0	18,0
13	17	18	-1	17,5
14	15	15	0	15,0
15	9	9	0	9,0
16	14	14	0	14,0
17	17	18	-1	17,5
18	18	18	0	18,0
19	20	20	0	20,0
20	10	11	-1	10,5

jetos). Para calcularlo necesitamos una serie de sujetos a los que se les realice al menos dos mediciones. En la tabla X se presentan los resultados de dos mediciones repetidas de bilirrubina transcutánea en recién nacidos ictéricos[42]. La desviación estándar intrasujetos puede calcularse fácilmente usando un programa que realice análisis de la varianza (ANOVA). En la tabla XI podemos ver la salida de ordenador del ANOVA para los datos de la tabla X. El parámetro denominado CMr (cuadrados medios de los residuos) es la varianza intrasujetos. Si realizamos la raíz cuadrada de CMr obtendremos la desviación estándar intrasujetos (s_i). La s_i puede calcularse igualmente a partir del ANOVA para estudios con más de 2 mediciones por sujeto.

Tabla XI Análisis de la varianza de una vía de los datos de la tabla X. CMp cuadrados medios de los pacientes. CMr cuadrados medios de los residuos

Fuente de variación	*Grados de libertad*	*Suma de cuadrados*	*Cuadrados medios*	
Pacientes	19	371,5000	19,5526	CMp
Residual	20	6,0000	0,3000	CMr
Total	39	377,5000		

Utilizando la s_i podemos cuantificar el margen de error de nuestras mediciones. Así, podemos estimar que la diferencia entre una medición determinada y el verdadero valor no será mayor de 1,96 veces la s_i en el 95% de las observaciones. También nos permite estimar que la diferencia entre dos mediciones repetidas en un mismo sujeto no superarán 2,77 veces la s_i en el 95% de las observaciones[43,44]. Para nuestro ejemplo, la s_i es 0,54, la diferencia estimada respecto al valor verdadero menor de 1,05 y la diferencia entre dos mediciones menor de 1,49.

Coeficiente de Correlación Intraclase

Si sólo se realizan dos mediciones por sujeto, la forma más intuitiva de compararlas es representarlas en un diagrama de puntos, examinar si existe relación lineal entre ambas y calcular su coeficiente de correlación. En la figura 2 se presenta el

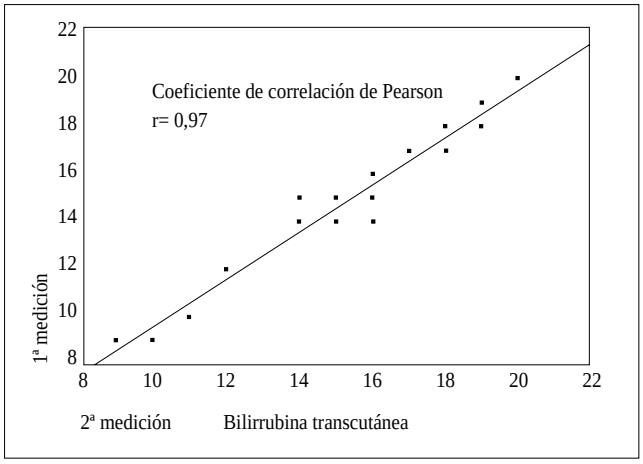

Figura 2. Diagrama de puntos y correlación lineal de los datos de la tabla X.

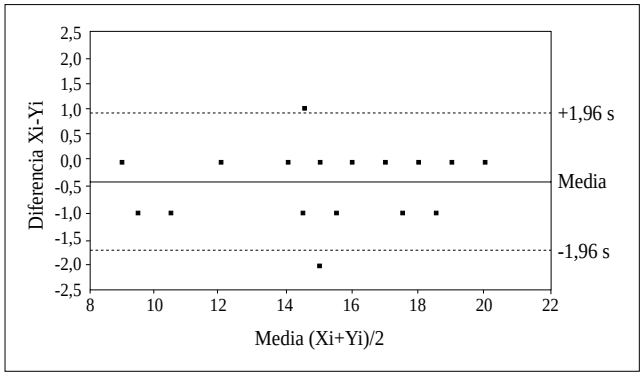

Figura 3. Método de Bland - Altman con los datos de la tabla X.

diagrama de puntos de los datos de la tabla X. El coeficiente de correlación de Pearson (r) para estos datos es 0,97.

Sin embargo, la existencia de una fuerte relación lineal con un alto coeficiente de correlación no indica que haya una buena concordancia entre las mediciones, solamente que los puntos del diagrama se ajustan a una recta. El coeficiente de correlación depende, en gran manera, de la variabilidad entre sujetos, por ello, varía mucho en función de las características de la muestra donde se estima, afectándole especialmente la presencia de valores extremos. Si una de las mediciones es sistemáticamente mayor que otra, el coeficiente de correlación será muy alto, a pesar de que las mediciones nunca concuerden. Estos problemas son evitados utilizando el coeficiente de correlación intraclase.

El coeficiente de correlación intraclase (CCI) estima la concordancia entre dos o más medidas repetidas. El cálculo del CCI se basa en un modelo de ANOVA con medidas repetidas, aplicándose distintas fórmulas en función del diseño y los objetivos del estudio[45]. El escenario más simple es aquél en el que estimamos la variabilidad de las medidas, sin tener en cuenta la variabilidad aportada por los distintos observadores (diseño de una vía con factor aleatorio). Considerando este diseño, y utilizando los resultados del ANOVA, podemos calcular el CCI con la siguiente fórmula:

$$CCI = \frac{CMp - CMr}{CMp + (k - 1)\,CMr}$$

donde k es el número de observaciones por sujeto, CMp son los cuadrados medios entre pacientes y CMr los cuadrados medios de los residuos.

Con los datos del ANOVA de la tabla XI el CCI será:

$$CCI = \frac{19{,}55 - 0{,}30}{19{,}55 + (2 - 1)\,0{,}30} = 0{,}996$$

En nuestro ejemplo, apenas hay diferencias entre el CCI y el coeficiente de correlación de Pearson (r). Si el CCI fuera mucho menor que r, habría que pensar que existe un cambio sistemático entre una medida y otra, lo que podría estar causado por un efecto de aprendizaje. En este caso, las mediciones no se han realizado en las mismas circunstancias, por lo que no se dan las condiciones para realizar un estudio de fiabilidad[46].

Método de Bland-Altman

Un método alternativo para analizar la concordancia entre 2 observaciones repetidas que se miden en una escala continua es el método gráfico descrito por Bland y Altman[47]. Consiste en representar en un diagrama de puntos la diferencia entre los pares de mediciones contra su media (Fig. 3). Ello permite examinar la magnitud de las diferencias y su relación con la magnitud de la medición. Además, se puede estimar la desviación estándar de las diferencias y los intervalos entre los que cabe esperar que se encuentre el 95% de las diferencias.

Cuando la variabilidad en las medidas no es constante, sino que cambia al aumentar o disminuir la magnitud de la medida, el cálculo se complica[48]. Si existe correlación significativa entre las diferencias y las medias, la variabilidad no será constante. En ese caso, puede intentarse realizar transformaciones logarítmicas de los datos o analizar la variabilidad por separado para varios intervalos de valores.

Bibliografía

1 Corral Corral C. El Razonamiento Médico. Ed. Díaz de Santos. Madrid 1994; 79-121.

2 Porta Serra M. La observación clínica y el razonamiento epidemiológico. *Med Clin* (Barc) 1986; 816-819.

3 Pozo Rodríguez F. La eficacia de las pruebas diagnósticas (I). *Med Clin* (Barc) 1988; **90**:779-785.

4 Hoberman A, Chao HP, Keller DM, Hickey R, Davis HW, Ellis D. Prevalence of urinary tract infection in febrile infants. *J Pediatr* 1993; **123**:17-23.

5 Hoberman A, Wald ER, Reynolds EA, Penchansky L, Charron M. Pyuria and bacteriuria in urine specimens obtained by catheter from young children with fever. *J Pediatr* 1994; **124**:513-519.

6 Muñoz C, Gené A, Pérez I, Mira J, Roca J, Latorre C. Diagnóstico de la tuberculosis en niños. Evaluación de la técnica reacción en cadena de la polimerasa. *An Esp Pediatr* 1997; **47**:353-356.

7 Bladé J, Alaman E, Cartaña A, Guinea I, Liberal A, Herreros M y col. Evaluación de los datos clínicos y de una técnica de detección rápida

(TestPack Stre A) en el diagnóstico de las faringoamigdalitis agudas estreptocócicas. *Atención Primaria* 1991; **8**:24-30.

8 Stevens JC, Webb HD, Smith MF, Buffin JT. Evaluation of click evoked OEA in the newborn. *Br J Audiol* 1991; **25**:11-14.

9 Maisels MJ, Kring E. Transcutaneous Bilirubinometry Decreases the Need for Serum Bilirubin Measurements and Saves Money. *Pediatrics* 1997; **99**:599-600.

10 Garrido Redondo M, Blanco Quirós A, Garrote Adrados JA, Tellería Orriols JJ, Arranz Sanz E. Valor de los anticuerpos salivales para la determinación de la seropositividad frente a sarampión, rubéola y parotiditis en niños y adultos. *An Esp Pediatr* 1997; **47**:499-504.

11 Guidelines for the diagnosis of rheumatic fever. Jones criteria, 1992 update. *JAMA* 1992; **268**:2069-2073.

12 Holm VA, Cassidy SB, Butler MG, Hanchett JM, Greenswag LR, Whitman BY, Greenberg F. Prader-Willi Syndrome: Consensus Diagnostic Criteria. *Pediatrics* 1993; **91**:398-402.

13 Phelps CE, Hutson A. Estimating Diagnostic tests accuracy using a Fuzy gold standard. *Med Decis Making* 1995; **15**:44-57.

14 Lohr JA, Portilla MG, Geuder TG, Dunn ML, Dudley SM. Making a presumptive diagnosis of urinary tract infection by using a urinalysis performed in an on-site laboratory. *J Pediatr* 1993; **122**:22-25.

15 Rossner B. Fundamentals of Biostatistics. Boston. PWS-KENT Publishing Company 1990; 170-175.

16 Coughlin SS, Picle LW. Sensitivity and specificity-like measures of the validity of a Diagnostic test that are corrected for chance agreement. *Epidemiology* 1992; **3**:178-181.

17 Rossner B. Fundamentals of Biostatistics. Boston. PWS-KENT Publishing Company 1990; 42-70.

18 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159-174.

19 Fleiss JL. The measurement of interrater agreement. En Fleiss JL, editor. Statistical Methods for Rates and Proportions. Toronto. John Wiley & Sons 1981; 212-236.

20 Rapkin RH. Urinary tract infection in childhood. *Pediatrics* 1977; **60**:508-511.

21 Wettergren B, Jodal U, Jonasson G. Epidemiology of bacteriuria during the first year of life. *Acta Paediatr Scand* 1985; **74**:925-933.

22 Bonilla Miera C, Rollán Rollán A, González de Aledo Linos A. Detección de alteraciones nefrourológicas en el lactante mediante tira reactiva, Experiencia en una consulta de puericultura. *An Esp Pediatr* 1988; **29**:244-247.

23 Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40**:513-517.

24 Jaffe DM, Fleisher GR. Temperature and Total White Blood Count as Indicators of Bacteriemia. *Pediatrics* 1991; **87**:670-674.

25 Moise A, Clément B, Ducimetière P, Bourassa MG. Comparison of Receiver Operating Curves Derived from the Same Population: A Bootstrapping Approach. *Comput Biomed Res* 1985; **18**:125-131.

26 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; **39**:561-577.

27 Guangqin MA, Hall WJ. Confidence Bands for Receiver Operating Characteristics Curves. *Med Decis Making* 1993; **13**:191-197.

28 Centor RM, Keightley GE. Receiver operating characteristic (ROC) curve area analysis using the ROC ANALYZER. SCAMC Proc 1989; 222-226.

29 Krieg AF, Abendroth TW, Bongiovanni MB. When is a diagnostic test result positive?. Decision tree models based on net utility and threshold. *Arch Pathol Lab Med* 1986; **110**:787-791.

30 Ransohof DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic test. *N Engl J Med* 1978; **299**:926-930.

31 Begg CB. Biases in the assessment of diagnostic test. *Stat Med* 1987; **6**:411-423.

32 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995; **1274**:645-651.

33 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 1994; **271**:389-391.

34 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What were the results and will they help me in caring for my patients? *JAMA* 1994; **271**:703-707.

35 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology-a basic science for clinical medicine. London: Little, Brown, 1991: 51-68.

36 Mant D. Testing a test: three critical steps. In: Jones R, Kinmonth A-L, eds. Critical reading for primary care. Oxford: Oxford University Press, 1995: 183-190.

37 Greenhalgh T. How to read a paper: Papers that report diagnostic or screening tests. *BMJ* 1997; **315**:540-543.

38 Feinstein AR. Diagnostic and spectral markers. En Feinstein AR, editor. Clinical Epidemiology. The architecture of clinical research. WB Saunders, 1985; 597-631.

39 Ochoa Sangrador C, Brezmes Valdivieso MF. Efectividad de los test diagnósticos. *An Esp Pediatr* 1995; **42**:473-475.

40 Rodríguez Artalejo F, Banegas Banegas JR, González Enríquez J, Martín Moreno JM, Villar Álvarez F. Análisis de decisiones clínicas. *Med Clin* (Barc) 1990; **94**:348-354.

41 Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *J Clin Epidemiol* 1992; **45**:1-7.

42 Ochoa C, Marugán V, Tesoro R, García MT, Hernández MT. Validez y Precisión de la Bilirrubina Transcutánea. Libro de Abstracts. XXVII Congreso de la Asociación Española de Pediatría. Oviedo 26 a 28 de Junio de 1997. *An Esp Pediatr* 1997; **99**:18.

43 Bland JM, Altman DG. Measurement error. *BMJ* 1996; **312**:1654.

44 Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ* 1996; **312**:1472.

45 Fleiss JL. The design and analysis of clinical experiments. Nueva York: John Wiley & Sons 1986: 1-32.

46 Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996; **313**:41-42.

47 Bland JM, Altman DG. Statistical methods for assessing agreement