

## Pediatría basada en la evidencia. Estudios de diagnóstico (1ª parte)

### *Evidence-based pediatrics. Diagnostic tests (1<sup>st</sup> part)*

Dra. Graciela Demirdjian<sup>a</sup>, Dra. Valeria Berlín<sup>a</sup> y Dr. Hernán Rowensztein<sup>a</sup>

#### RESUMEN

Arribar a un diagnóstico correcto es un gran desafío que enfrenta el médico diariamente. A través de la anamnesis y del examen físico se llega a la sospecha clínica y con frecuencia se requiere un estudio auxiliar para confirmarla o descartarla. El avance de la tecnología ha hecho que el número de estudios disponibles para una misma enfermedad sea cada vez mayor. El médico debe elegir cuál es el adecuado, según el desempeño de la prueba, características del paciente, riesgos y costos. En este artículo analizaremos la confiabilidad, interpretación y aplicación de los resultados de una prueba diagnóstica.

**Palabras clave:** diagnóstico, medicina basada en la evidencia, técnicas y procedimientos de laboratorio, estudio de validación.

#### SUMMARY

To arrive to an accurate diagnosis is a great cha-

llenge in medical daily practice. Through anamnesis and physical examination, it is possible to have a clinical suspicion but it is often required a diagnosis test to confirm or rule out it. For the same condition, technological advances have multiplied the available tests. From them, doctors have to choose the most suitable test, according to its performance, patient characteristics, harms and costs. We will analyze the reliability, interpretation and application for one diagnosis test results.

**Key words:** diagnosis, evidence-based medicine, laboratory techniques and procedures, validation study

#### INTRODUCCIÓN

En este espacio dedicado a la Pediatría basada en la evidencia nos

#### Hinojosa-Pérez JO y col. Utilidad de la gasometría capilar para detectar hiperoxemia en el recién nacido grave. *Bol Med Hosp Infant Mex* 1999; 56(2): 93-96.

El monitoreo de oxígeno en el recién nacido (RN) es de vital importancia para corregir estados de baja concentración de este gas y evitar un exceso en sangre que puede provocar daño irreversible de la retina. La gasometría capilar ha mostrado buena correlación para pH y PcapO<sub>2</sub>, no así para pO<sub>2</sub>. Sin embargo, se continúa recurriendo a ella cuando no se dispone de un catéter arterial, tomando como base valores de presión parcial de oxígeno que se establecieron como normales en los años '60 (35-50 mmHg).

**Objetivo:** Determinar la sensibilidad y especificidad de la gasometría capilar para detectar hiperoxemia en el recién nacido grave.

**Material y métodos:** Estudio observacional, transversal, prospectivo, de tipo prueba diagnóstica, realizado en la unidad de cuidados intensivos neonatales de un hospital regional mexicano de junio a octubre de 1996. Se incluyeron RN con dificultad respiratoria grave, en ventilación mecánica, en los que fue posible instalar un catéter arterial. Se excluyeron del estudio RN en estado de shock o con mala perfusión tisular, hipotermia o hipertermia. Se tomaron 100 muestras para gasometrías en forma simultánea de catéter arterial y de punción capilar del talón, utilizando un gasómetro; efectuando primero la gasometría arterial y después la capilar. Previa a la toma de muestras de sangre se midió la saturación de oxígeno-hemoglobina en un aparato pulsoximétrico. El catéter arterial se colocó postductal y la toma de capilar se efectuó con la técnica de Blumenfeld. Se consideró como hiperoxemia una presión arterial de O<sub>2</sub> (PaO<sub>2</sub>) > 81 mmHg con una saturación de oxígeno-hemoglobina >95%. Los resultados se analizaron estadísticamente con la prueba de  $\chi^2$ . Se efectuó un análisis de sensibilidad, especificidad y valores predictivos de la prueba (gasometría capilar) para detectar hiperoxemia, realizando cortes cada 5 mmHg a partir de 40 mmHg. Además se analizó la razón de probabilidad y la probabilidad después de la prueba para cada punto de corte.

**Resultados:** De 48 RN que presentaron hiperoxemia por el gas arterial, 37 tuvieron una presión capilar de O<sub>2</sub> (PcapO<sub>2</sub> o PcapO<sub>2</sub>) >50 mmHg, en comparación a 11 RN de los 52 con valores arteriales normales (p<0,01). La sensibilidad de la prueba para detectar hiperoxemia con una PcapO<sub>2</sub> mayor de 55 mmHg fue de 77% y la especificidad de 78%; el valor predictivo positivo fue de 77% y el valor predictivo negativo de 78%. Tomando en cuenta valores de PcapO<sub>2</sub> mayores de 50 mmHg, la razón de probabilidad fue de 3,6 y la probabilidad posprueba de 78%.

**Conclusiones:** Los valores considerados normales de PcapO<sub>2</sub> tienen baja sensibilidad y especificidad para detectar hiperoxemia en el RN grave.

a. Hospital de  
Pediatría "Dr. Prof.  
Juan P. Garrahan".

*Conflicto de intereses:*  
Nada que declarar.

*Correspondencia:*  
Dra. Graciela Demirdjian:  
gdemir@intramed.net

Recibido: 3-2-09  
Aceptado: 19-6-09

Guía del usuario para artículos sobre diagnóstico. *JAMA* 1994; 271: 389-391, 703-707.

### Validez interna

#### 1. Criterios primarios:

- (a) ¿Existió una comparación ciega e independiente con un estándar de referencia?
- (b) ¿Se evaluó la prueba en un espectro apropiado de pacientes similares a aquellos en los que se aplicará en la práctica?

#### 2. Criterios secundarios:

- (c) ¿Los resultados de la prueba en estudio influyeron en la decisión de realizar el método de referencia ("gold standard")?
- (d) ¿Se describieron los métodos para llevar a cabo el examen con el suficiente detalle como para permitir su reproducción?

### Análisis de resultados

- (a) ¿Cuáles son los coeficientes de probabilidad (LR) asociados a diferentes intervalos de resultados de la prueba?
- (b) ¿Cuál es la precisión de las estimaciones?

### Utilidad de los resultados para la asistencia de mis pacientes

- (a) ¿Los resultados son aplicables a los pacientes en mi práctica?
- (b) ¿Será posible realizar e interpretar la prueba en forma satisfactoria en mi medio?
- (c) ¿Los resultados van a cambiar mi estrategia de tratamiento?
- (d) ¿Los pacientes van a estar mejor como resultado de la prueba?

ocupamos hoy de los estudios de validación de pruebas diagnósticas. La lectura crítica de estos estudios es fundamental para la toma de decisiones respecto de la solicitud de estudios diagnósticos y la interpretación de sus resultados.

Como se trata de un tema algo complejo lo hemos dividido en dos artículos. En la primera parte de esta serie nos abocaremos al análisis de las pruebas diagnósticas con resultado dicotómico (positivo o negativo). En la segunda parte abordaremos las pruebas con resultado numérico continuo y el análisis de los puntos de corte.

Para esta oportunidad hemos elegido un artículo publicado en una revista latinoamericana sobre un tema simple y muy conocido por los pediatras: las muestras de gases capilares (el recuadro aporta un resumen, pero para la lectura crítica sugerimos utilizar el texto completo del artículo recurriendo a la fuente citada).<sup>1</sup>

### Diseño de estudios de validación de pruebas o tests diagnósticos

El diseño óptimo para validar un método diagnóstico es comparativo, transversal, observacional y prospectivo:

- Comparativo: se necesita comparar la capacidad diagnóstica del método en cuestión contra un test de referencia.<sup>a</sup>
- Transversal: cada prueba se realiza una sola vez en cada paciente.
- Observacional: el investigador observa sin realizar ninguna intervención.

- Prospectivo: las mediciones deben planificarse cuidadosamente, estandarizarse y hacerse en las condiciones correctas para garantizar su confiabilidad.

El estudio que nos ocupa especifica su diseño como observacional, transversal y prospectivo, (algunas otras características del diseño las analizaremos entre los criterios de validez interna).

### Cómo iniciar la lectura crítica

Para estar seguros de que comprendemos el objetivo del estudio, siempre es conveniente identificar sus cuatro componentes ("PICO"): el paciente (población), la intervención (la prueba diagnóstica), el comparador (el test de referencia) y el "outcome" o resultado (el evento o enfermedad que se quiere diagnosticar).

En este caso se compara en recién nacidos graves en asistencia respiratoria (pacientes) la gasometría capilar (intervención diagnóstica) contra la arterial (comparador) para detectar hiperoxemia (resultado o outcome).

---

a. A diferencia de los estudios de intervención donde la comparación se efectúa entre datos de dos conjuntos de sujetos (comparación intergrupo o intersujeto), en este caso la comparación se denomina intrasujeto o intragrupo porque el resultado de la prueba en cada paciente se compara contra el resultado del patrón de referencia en el mismo sujeto.

## ¿Sigo leyendo?

Antes de analizar los resultados y aplicarlos a nuestro paciente, es importante establecer si el estudio seleccionado se realizó correctamente y si por ende sus conclusiones son confiables. Las guías de usuarios publicadas en *JAMA* proponen como siempre tres secciones básicas: la validez interna del estudio, la magnitud y precisión de los resultados y su aplicabilidad o validez externa (ver recuadro).

## GUÍA PARA EL ANÁLISIS CRÍTICO:<sup>2</sup>

### A. Validez interna: ¿Son válidos los resultados del estudio?

En esta sección se valora si el estudio se diseñó adecuadamente y si carece de sesgos o errores sistemáticos que pudieran viciar sus conclusiones.

1. **Criterios primarios:** son aquellos aspectos que, de no cumplirse, generan sesgos que invalidan la investigación.

(a) ¿Existió una comparación ciega e independiente con un estándar de referencia?

Lo primero que evalúa esta pregunta es la elección del comparador. Para evaluar la utilidad de un método diagnóstico, éste debe ser comparado contra el mejor estudio conocido para reconocer la enfermedad o evento en cuestión. Este patrón o estándar de oro (*gold standard*) debe elegirse criteriosamente, ya que al ser la prueba que permite acercarse más a la certeza diagnóstica, permitirá valorar la utilidad del nuevo método en función de la concordancia de sus resultados. Habitualmente los estándares de referencia son estudios costosos, cruentos o difíciles de realizar o interpretar (biopsias, cultivos, imágenes sofisticadas), por lo que el objetivo principal de este tipo de investigaciones es "validar" un nuevo método (ver si éste se acerca a la eficacia diagnóstica del patrón o estándar) que tenga ventajas en cuanto a su sencillez, rapidez, riesgos o costos.

En segundo término se valora si las mediciones de ambas pruebas se realizaron de manera ciega o enmascarada. Nuevamente nos encontramos con que el cegamiento o enmascaramiento se aplica para controlar la subjetividad en las mediciones: es importante que quien evalúa el resultado de un método diagnóstico no conozca el resultado de la otra prueba para evitar la interpretación subjetiva del observador. La importancia de un enfoque a ciegas es mayor cuanto mayor sea la influencia de cono-

cer los resultados previos, sobre todo si se empieza por el estándar (por ejemplo: si se efectúa una tomografía computada para el diagnóstico de bronquiectasias es probable que una vez conocido el resultado de ésta se observen imágenes sospechosas en la radiografía de tórax que, quizás, en otro caso no hubieran llamado la atención).

Finalmente se requiere que ambas mediciones sean independientes, es decir, que ambos tests se apliquen en todos los pacientes, independientemente de que el primer resultado sea positivo o negativo.

En este artículo se compara un gas capilar extraído del talón contra una muestra arterial obtenida de un catéter umbilical como "estándar de oro". Ambas muestras se tomaron en todos los RN de manera independiente. No se aclara si las muestras fueron interpretadas por distintos observadores o en forma ciega, aunque aquí esto es irrelevante pues los resultados son numéricos y no influenciados por la subjetividad.

(b) ¿Se evaluó la prueba en un espectro apropiado de pacientes similares a aquellos en los que se aplicará en la práctica?

En este punto se evalúa el tipo de pacientes incluidos en el estudio. Está claro que ante un caso grave o florido de enfermedad cualquiera hace el diagnóstico... Esto también ocurre con los estudios diagnósticos. Para que la validación de la nueva prueba no sobrestime o subestime su capacidad diagnóstica, el estudio debe incluir un amplio espectro de pacientes que incluya sanos y enfermos en toda su gama, y el ámbito del estudio debe ser similar al ámbito de aplicación de la prueba diagnóstica. De no ser así, se genera un "sesgo de espectro": un método diagnóstico validado en un centro de referencia de una especialidad suele sobrestimar su capacidad diagnóstica respecto de su aplicación en un centro de atención primaria o en la población general).

En los criterios de inclusión del artículo se observa que los pacientes elegibles fueron aquellos RN a los que habitualmente se solicitaría una muestra de gases en sangre para descartar hiperoxemia (neonatos con dificultad respiratoria grave en asistencia ventilatoria mecánica), por lo que el ámbito de estudio y de aplicación coinciden. Sin embargo, sólo se incluyó a aquellos en los que fue posible colocar un catéter arterial,

lo que podría haber generado un sesgo de espectro al excluir pacientes más graves o problemáticos. También se excluyeron pacientes con shock o alteraciones de la perfusión tisular o de la temperatura corporal, situaciones que atentan contra la confiabilidad de un gas capilar. Todos estos criterios de exclusión son razonables para aumentar la validez interna del estudio, pero es importante notar que también limitan su aplicabilidad (validez externa) a pacientes con similares características.

2. **Criterios secundarios:** son características metodológicas que implican un valor agregado a la validez del estudio.

(c) ¿Los resultados de la prueba en estudio influyeron en la decisión de realizar el estándar de referencia?

Este es un aspecto en el que el contexto de investigación se diferencia del asistencial. Habitualmente, en la atención de un paciente el proceso diagnóstico se inicia con pruebas más rápidas y sencillas, y se reservan las más cruentas o costosas para aquellos con resultados positivos. Si esto mismo se hiciera en un estudio de validación de una nueva prueba, sólo los pacientes con test positivo (los más graves) llegarían a realizarse la prueba de referencia, lo cual genera lo que se denomina "sesgo de verificación".

En nuestro ejemplo no existe sesgo de verificación dado que el estándar de oro (muestra arterial) fue realizado en todos los pacientes con hiperoxemia o sin ella en el gas capilar.

(d) ¿Se describieron los métodos para llevar a cabo el examen con el suficiente detalle como para permitir su reproducción?

Describir detalladamente la metodología es fundamental para asegurar la replicabilidad de cualquier investigación. En una validación de métodos diagnósticos, la descripción de las condiciones en que se efectuaron las mediciones (incluidos la preparación del paciente, la calibración de los instrumentos,

el entrenamiento de los observadores y la estandarización de los procedimientos) es de vital trascendencia, ya que el estudio se centra, precisamente, en valorar los resultados de estas mediciones.

En el apartado de Materiales y Métodos del trabajo mencionado se describen detalladamente las condiciones en las que se realizaron ambas pruebas, incluidos la técnica de colocación del catéter arterial, el calentamiento del talón previo a la toma de la muestra capilar y la definición operativa de hiperoxemia.

### B. Análisis de resultados: ¿Son los resultados importantes?

(a) ¿Cuáles son los coeficientes de probabilidad (LR) asociados a diferentes intervalos de resultados de la prueba?

(b) ¿Cuál es la precisión de las estimaciones?

Los resultados de los estudios diagnósticos pueden presentarse en dos tipos de valores: dicotómicos (prueba positiva-prueba negativa) o continuos (datos numéricos como  $pO_2$ , glucemia, hematócrito, número de colonias en un cultivo). En esta primera entrega sólo abordaremos los estudios diagnósticos con resultados dicotómicos (pero recordemos que los siguientes criterios también son aplicables a cualquier prueba con resultado numérico que se ha "dicotomizado" eligiendo un valor límite o punto de corte determinado).

Datos dicotómicos: en este caso se analiza en qué grado el método diagnóstico permite distinguir entre enfermos y sanos. Esta información se resume con las denominadas medidas de capacidad operativa: sensibilidad, especificidad, poder predictivo positivo y poder predictivo negativo. Para calcularlas es preciso construir una tabla de  $2 \times 2$  (Tabla 1) donde se distribuyan los resultados obtenidos de ambos estudios: el test a prueba y el estándar de oro. En ella se observa que, dado que el estándar se utiliza como criterio de verdad, cuando sea positivo, el paciente se definirá como enfermo y cuando sea negativo se considerará sano (estos datos se ubican en las columnas). Los

TABLA 1. Tabla de  $2 \times 2$  para el cálculo de la capacidad operativa de una prueba diagnóstica

	Prueba de referencia positiva	Prueba de referencia negativa	
Prueba positiva	Verdaderos Positivos (VP)	Falsos Positivos (FP)	Positivos
Prueba negativa	Falsos Negativos (FN)	Verdaderos Negativos (VN)	Negativos
	Enfermos	Sanos	

resultados del método en estudio pueden coincidir (resultados verdaderos) o no (resultados falsos) con el test de referencia (estos datos se ubican en las filas).

Veamos ahora cómo se calculan y qué significan estas medidas de capacidad operativa:<sup>1</sup>

- Sensibilidad (S) o tasa de verdaderos positivos (TVP): La sensibilidad de un método diagnóstico es la capacidad de identificar a los pacientes enfermos. Indica la proporción de pacientes enfermos a los que el test les dio positivo, y se expresa como porcentaje (0-100%) o fracción (de 0 a 1). Una prueba con alta sensibilidad tiene pocos falsos negativos (pacientes enfermos a los que la prueba no detectó). (Al observar la tabla 2 x 2 se nota que para este cálculo se utiliza la columna de enfermos que es el denominador de la sensibilidad).

**Sensibilidad= Verdaderos Positivos/ Enfermos= VP/ (VP + FN)**

(FN: falsos negativos)

- Especificidad (E) o Tasa de Verdaderos Negativos (TVN): La especificidad es la capacidad de un método diagnóstico de reconocer a las personas sanas. Una especificidad alta evita falsos positivos (pacientes sanos a los que el test cataloga como positivos). (Como el denominador de este cálculo son los sanos se utilizan los datos de la segunda columna).

**Especificidad= Verdaderos Negativos/ Sanos= VN/ (VN + FP)**

(FP: falsos positivos)

- Poder Predictivo Positivo (PPP) o Valor Predictivo Positivo (VPP): El poder predictivo positivo mide la probabilidad de estar enfermo si la prueba es positiva. Es la proporción de pruebas positivas que son verdaderas o corresponden a pacientes enfermos. Una prueba con alto poder predictivo positivo tiene pocos falsos positivos (resultados positivos equivocados). (En la tabla de 2 x 2 las pruebas con resultado positivo se ubican en la primera fila, la de las pruebas positivas, que es la que se utiliza para este cálculo).

**PPP= Verdaderos positivos/ Positivos= VP/ (VP + FP)**

- Poder Predictivo Negativo (PPN) o Valor Predictivo Negativo (VPN): El poder predictivo negativo mide la probabilidad que tiene un paciente de estar sano si la prueba le dio negativa. Es la proporción de resultados negativos ver-

daderos o correspondientes a pacientes sanos. Una prueba con alto poder predictivo negativo tiene pocos falsos negativos (resultados negativos erróneos). (Para este cálculo se utiliza la segunda fila de la tabla de 2 x 2 que corresponde a las pruebas con resultado negativo).

**PPN= Verdaderos Negativos/ Negativos= VN/ (VN + FN)**

Estas medidas de capacidad operativa tienen algunas limitaciones:

- Los denominadores de sensibilidad y especificidad son los enfermos y sanos respectivamente, por lo que no nos sirven para aplicar a un paciente particular, ya que cuando solicitamos un estudio lo hacemos precisamente porque no conocemos su estado de salud o enfermedad. Conocer la sensibilidad y especificidad de una prueba sólo nos ayudan al momento de elegir cuál solicitar:
  - Una prueba muy sensible es útil al inicio del proceso diagnóstico, para pesquisa o rastreo (*screening*), para descartar enfermedad y para enfermedades en las que el riesgo es no tratar (porque tiene pocos FN).
  - Una prueba muy específica es útil al final del proceso diagnóstico, para confirmación diagnóstica de enfermedad y para situaciones en las que el riesgo es "etiquetar" como enfermo o tratar de más (porque tiene pocos FP).
- Los valores predictivos, en cambio, son aplicables al paciente individual porque lo que sí conocemos del paciente es si la prueba dio positiva o negativa. El poder predictivo nos dirá qué probabilidad tiene este resultado de ser verdadero y, por ende, qué importancia diagnóstica tendrá. Sin embargo, estas medidas tienen la desventaja de que varían con la prevalencia de enfermedad:
  - Una prevalencia alta aumenta el PPP de una prueba (porque pocos resultados positivos serán FP).
  - Una prevalencia baja aumenta el PPN de una prueba (porque pocos resultados negativos serán FN).
- Ninguna de estas medidas utiliza los datos de toda la tabla, por lo que la información que transmiten es parcial (habría que recordar las cuatro medidas para el desempeño de una prueba determinada).  
Existe otra medida de capacidad operativa que supera estas limitaciones, al combinar, en una



única cifra, la información de toda la tabla y que es aplicable al caso particular: el cociente de probabilidades o razón de probabilidades.

- Razón de probabilidades o coeficiente de verosimilitudes (o "Likelihood Ratio": LR): Expresa la chance ("odds"<sup>b</sup>) de estar enfermo cuando la prueba es positiva (LR positivo) o negativa (LR negativo). Compara cuántas veces la prueba acierta contra las veces que se equivoca:

**LR+:** Tasa de Verdaderos Positivos/ Tasa de Falsos Positivos = TVP/ TFP = S/ 1-E

**LR-:** Tasa de Falsos Negativos/ Tasa de Verdaderos Negativos = TFN/ TVN = 1-S/ E

Un LR= 1 significa que las chances de un paciente de estar enfermo o sano son iguales (la prueba no ayuda al diagnóstico para la patología en estudio). Para que un método diagnóstico tenga utilidad se espera que su LR+ sea alto (idealmente > 5-10) y su LR- sea bajo (< 0,1-0,2).

Calculemos estas medidas de desempeño o capacidad diagnóstica para un punto de corte (valor límite) del gas capilar de 50 mmHg con los datos del artículo (Tabla 2).

- $S = VP / VP + FN = VP / ENFERMOS = 37 / 37 + 11 = 37 / 48 = 0,77 = 77\%$  (IC 95%: 64%-87%)
- $E = VN / VN + FP = VN / SANOS = 41 / 41 + 11 = 41 / 52 = 0,798 = 79\%$  (IC 95%: 66%-88%)
- $PPP = VP / VP + FP = VP / POSITIVOS = 37 / 37 + 11 = 37 / 48 = 0,77 = 77\%$  (IC 95%: 64%-87%)
- $PPN = VN / VN + FN = VN / NEGATIVOS = 41 / 41 + 11 = 41 / 52 = 0,79 = 79\%$  (IC 95%: 66%-88%)
- $LR + = SENSIBILIDAD / 1 - ESPECIFICIDAD = TVP / TFP = 0,77 / 1 - 0,78 = 0,77 / 0,22 = 3,6$  (IC 95%: 2-6)

- $LR - = 1 - SENSIBILIDAD / ESPECIFICIDAD = TFN / TVN = 1 - 0,77 / 1 / 0,78 = 0,23 / 0,78 = 0,3$  (IC 95%: 0,2-0,5)<sup>c</sup>

### ¿Cómo interpretamos estos números?

- Una sensibilidad de 77% significa que, de todos los pacientes enfermos (con hiperoxemia), la prueba reconoce como positivos el 77% (TVP) y se pierde de diagnosticar el 23% (TFN).
- Una especificidad de 78% significa que, de todos los pacientes sanos (sin hiperoxemia), la prueba reconoce como negativos el 78% (TVN) y cataloga erróneamente como positivos el 22% (TFP).
- Un PPP de 77% implica que, de todas las muestras de gas capilar positivas, serán VP el 77% y FP el 23%.
- Un PPN de 78% implica que, de todas las muestras de gas capilar negativas, serán VN el 78% y FN el 22%.
- Un LR+ de 3,5 indica que un paciente con gas capilar positivo (> 50 mmHg) tiene 3 veces y media más chance de estar enfermo (hiperoxémico) que sano.
- Un LR- de 0,29 indica que la chance de estar enfermo (hiperoxémico) de un paciente con gas capilar negativo ( $\leq$  50 mmHg) es de 0,3 a 1. Veamos ahora cómo se puede aplicar todo esto al caso individual.
- Probabilidad preprueba y posprueba: Cuando se solicita una prueba diagnóstica se supone que el paciente presenta ciertas características que hacen sospechar la enfermedad en estudio (antecedentes, sintomatología, edad o factores de riesgo). A esta probabilidad de que el paciente esté enfermo antes de confirmarlo con un método diagnóstico se la denomina proba-

TABLA 2. Capacidad operativa de la gasometría capilar tomando como punto de corte una presión arterial de O<sub>2</sub> de 50 mmHg (Cuadro 1 en el artículo original)

Prueba diagnóstica (muestra capilar)	Prueba de referencia (muestra arterial)		
	Enfermos (CON hiperoxemia)	Sanos (SIN hiperoxemia)	
Resultado positivo	Verdaderos positivos (VP) 37	Falsos positivos (FP) 11	Total de positivos 48
Resultado negativo	Falsos negativos (FN) 11	Verdaderos negativos (VN) 41	Total de negativos 52
	Total de enfermos 48	Total de sanos 52	Total de muestras 100

b. Una razón ("ratio", "odds" o "chance") es un cociente entre dos parciales, mientras que una proporción (o probabilidad) es un cociente entre un parcial y el total.

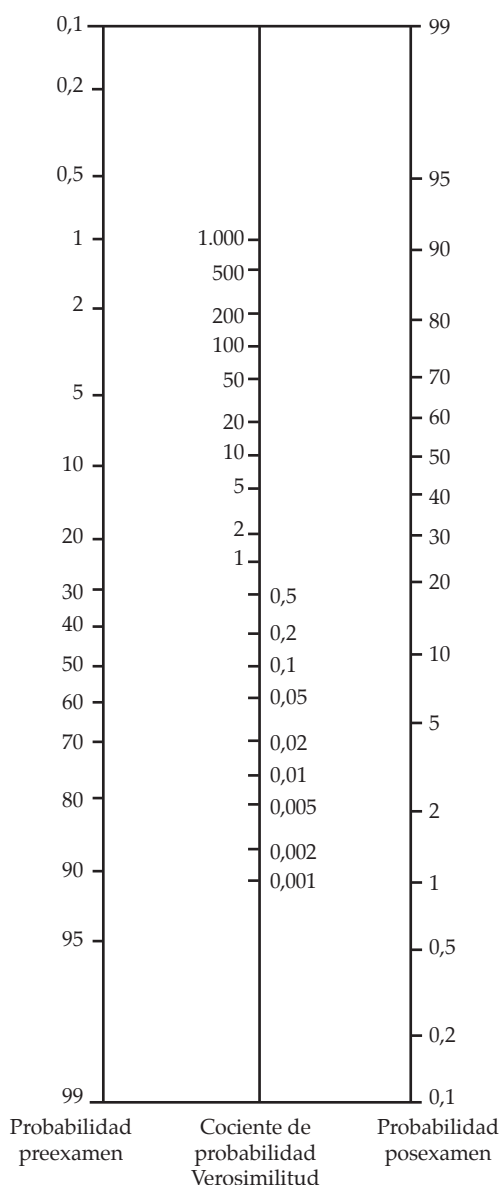
c. Como siempre, todas estas medidas estadísticas son estimaciones y su precisión está expresada por el intervalo de confianza. Como los autores del trabajo no los informan, los calculamos para estos datos y los agregamos.

bilidad preprueba. Se estima en base a los datos del paciente y el conocimiento (experiencia médica, bibliografía, trabajos científicos) de la enfermedad y su prevalencia (por ejemplo: la prevalencia de enfermedad celíaca entre familiares de primer grado es del 10%; en un paciente de 18 meses que tiene el mismo peso desde los 9 meses y presenta un abdomen globoso y nalgas fundidas se podría estimar una probabilidad de enfermedad celíaca del 80-85%). Cuando no se conoce ningún dato del paciente, la probabilidad preprueba equi-

vale a la prevalencia de la enfermedad en la población. El resultado de la prueba aumentará o disminuirá esta probabilidad basal, acercando o alejando el diagnóstico presuntivo, y transformándose en probabilidad posprueba. El impacto de la prueba diagnóstica para mover de probabilidad preprueba a posprueba se mide con el valor del LR:

- Un LR > 1 significa que la prueba aumenta la probabilidad de enfermedad, es decir que la probabilidad posprueba es mayor que la preprueba (a mayor LR, mayor utilidad de la prue-

FIGURA 1. Nomograma de Fagan para cálculo de probabilidad posprueba<sup>4,5</sup>



ba para arribar al diagnóstico sospechado).

- Un LR < 1 significa que la prueba disminuye la probabilidad de enfermedad por debajo de la que tenía antes de realizar el estudio, alejando el diagnóstico (a menor LR, mayor utilidad de la prueba para alejar el diagnóstico).

Como las probabilidades son proporciones y el LR es una razón, no se pueden multiplicar directamente, por lo que para obtener la probabilidad posprueba son necesarias algunas transformaciones matemáticas algo engorrosas. Por suerte, existe otra forma más sencilla de calcular la probabilidad posprueba utilizando un nomograma (Figura 1).

Una vez estimada la probabilidad preprueba, se traza una línea que pase por ésta y el valor del LR y se obtiene así la probabilidad posprueba.

En nuestro artículo, la probabilidad preprueba es la prevalencia de hiperoxemia en la muestra (48%).

- Si la prueba es positiva (gas capilar con  $pO_2 > 50$  mmHg) su LR+ es de 3,5 y la probabilidad posprueba sube a 77%.
- Si la prueba es negativa (gas capilar con  $pO_2 \leq 50$  mmHg) su LR+ es de 0,3 y la probabilidad posprueba baja a 22%.

### C. Validez externa: ¿Me ayudarán los resultados en la asistencia de mis pacientes?

Finalmente, antes de decidir usar la prueba diagnóstica, es necesario analizar algunas condiciones de aplicabilidad de los resultados al propio contexto.

- (a) ¿Son aplicables los resultados a mi paciente?

Lo primero es evaluar si el estudio se realizó en un contexto similar al de su práctica. Si las poblaciones son muy diferentes en cuanto a criterios de inclusión y exclusión, el desempeño de la prueba observado en el estudio puede no ser directamente trasladable a su paciente.

- (b) ¿Está la prueba disponible, es accesible, reproducible y fácilmente interpretable?

Es fundamental que la prueba diagnóstica pueda realizarse en su práctica diaria, es decir, que se tenga acceso a ella, que sea costeable, de bajo riesgo (poco invasiva) y que el desempeño sea constante. Esto último involucra ya al operador de la prueba, si ésta es subjetiva, o requiere una cierta experiencia en el observador.

- (c) ¿Los resultados modificarán mi tratamiento?

El método diagnóstico será de máxima utilidad en los pacientes con una probabilidad preprueba intermedia (zona de mayor incertidumbre en cuanto al estado de salud o enfer-

medad) y en los que el valor del *likelihood ratio* o coeficiente de probabilidad esté más alejado de 1, ya que en estos casos se observará el mayor cambio de probabilidad preprueba a posprueba y, por ende, la mayor influencia sobre la toma de conducta.

- (d) ¿Obtendrán beneficio los pacientes como consecuencia del examen?

El verdadero impacto de un test ocurre cuando, a partir del resultado, el diagnóstico se hace con más precisión o más precozmente, conduciendo a un tratamiento efectivo y a mejores resultados finales. Este aspecto sólo puede valorarse de manera óptima en estudios que evalúan la prueba diagnóstica como una intervención (un ensayo clínico controlado y aleatorizado). Los estudios de validación que hemos estado tratando sólo evalúan rendimiento diagnóstico y no garantizan impacto sobre la salud de los pacientes.

Somos conscientes de que éste es un tema complicado... Nuestro proceso diagnóstico habitual parece ser más una cuestión intuitiva y de "ojo clínico" que un problema epidemiológico. Pero también sabemos que la aplicación de los conceptos de esta guía puede optimizar la tarea diagnóstica y promover un uso más adecuado de los estudios disponibles. Como siempre, la incorporación de las herramientas de la Medicina Basada en la Evidencia debe hacerse de manera comprensiva, gradual y a la luz de la experiencia clínica. Sugerimos empezar por conocer los valores de capacidad diagnóstica de los estudios que solicitamos más habitualmente a nuestros pacientes; esta información es cada vez más fácil de encontrar en la bibliografía disponible y nos permitirá seleccionar la prueba más eficiente para la situación específica. Para completar y profundizar la información sintetizada en este artículo, incluimos en la bibliografía algunas fuentes adicionales de lecturas sugeridas, varias en castellano.<sup>6-10</sup> Continuaremos con este tema en la próxima entrega, en la que analizaremos las pruebas con resultado numérico continuo, los puntos de corte y las curvas ROC. ■

### BIBLIOGRAFÍA

1. Hinojosa-Pérez JO, Trevino-Baez JD. Utilidad de la gasometría capilar para detectar hiperoxemia en el recién nacido grave. *Bol Méd Hosp Infant Mex* 1999;56(2):93-96.
2. Jaeschke R, Guyatt GH, Sackett DL. Guía para usuarios de la literatura médica. Cómo utilizar un artículo sobre un examen diagnóstico. *JAMA* 1994;271:389-392 y 703-707.
3. Demirdjian G. Estudios de validación de métodos diagnósticos. En: Programa de Educación a distancia en Me-



- Metodología de la Investigación para Pediatría (PREMIP). Nivel 2. Módulo 2.8, 2003.
4. Faggan TJ. Nomogram for Bayes theorem. *NEJM* 1975;293:257.
  5. Ochoa Sangrador C, González de Dios J, Buñuel Álvarez JC. Evaluación de artículos científicos sobre pruebas diagnósticas. *Evid Pediatr* 2007;3:24.
  6. Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. *BMJ* 1997;315:540-543.
  7. López-Giménez F, Rohde LE, Luna-Giménez MA. Problemas y soluciones en la interpretación de pruebas diagnósticas. *Rev Investig Clín* 1998;50:65-72.
  8. Abraira V. Sesgos en los estudios de pruebas diagnósticas. *SEMERGEN* 2006;32(1):24-26.
  9. Abraira V. Índices de rendimiento de las pruebas diagnósticas. *SEMERGEN* 2002;28(4):193-194.
  10. García-García JJ. Significado y empleo de la razón de probabilidades en la práctica clínica. *Rev Mex Ped* 2000;67(4):188-191.

La vida es una obra de teatro que no permite ensayos. Por eso, canta, llora, baila, ríe y disfruta intensamente cada momento de tu vida, antes de que el telón baje y la obra termine sin aplausos.

*Charles Chaplin*