Pediatría basada en la evidencia. Estudios de diagnóstico (2ª parte)

Evidence-based pediatrics. Diagnostic tests (2nd part)

Dra. Graciela Demirdjian[a]

RESUMEN

El proceso diagnóstico es complejo y ejercitarlo exige experiencia e instrumentos adecuados. El avance tecnológico ha multiplicado el número de pruebas diagnósticas disponibles, aunque son pocas las herramientas clínicas bien diseñadas, validadas y útiles. En este artículo describimos el proceso de construcción de reglas de predicción clínica y el uso de curvas ROC para la selección del valor límite óptimo para una prueba con resultado numérico.

Palabras clave: *medicina basada en la evidencia, diagnóstico, pronóstico, curva ROC, reglas de predicción clínica.*

SUMMARY

Diagnosis is a complex process, demanding experience and proper instruments. Technology has advanced rapidly, increasing the number of available diagnostic tests. However, few well designed and validated useful clinical tools exist. This article summarizes the development process for clinical prediction rules and the use of ROC curves to select the best cutoff point for tests with continuous results.

Key words: *evidence-based medicine, diagnosis, prognosis, ROC curve, clinical prediction rules.*

INTRODUCCIÓN

En la primera parte de este artículo, publicada en la Sección de Pediatría basada en la evidencia, iniciamos el análisis crítico de los estudios de validación de pruebas diagnósticas[1] según las Guías del JAMA[2] y utilizamos, como ejemplo, un artículo sobre gases capilares y arteriales para el diagnóstico de hiperoxemia en neonatos.[3] En esta segunda parte abordaremos dos temas conexos algo más complejos: las curvas ROC y las reglas de predicción clínica.

CURVAS ROC

Nuestro análisis crítico del artículo sobre diagnóstico de hiperoxemia en neonatos había considerado la capacidad operativa de la prueba tomando sus resultados en formato dicotómico. Esto presupone que la prueba sólo puede ofrecer un resultado cualitativo (positivo-negativo), o bien que los resultados se expresan en una escala numérica continua, pero se conoce cuál es el valor límite para discriminar entre enfermos y sanos (el punto de corte o "*cut-off point*"). Seleccionar un punto de corte óptimo puede parecer sencillo, pero en realidad involucra una serie de consideraciones acerca de cómo se utilizará la prueba diagnóstica.

Para empezar, recordemos que la sensibilidad y especificidad sirven principalmente para elegir la prueba a utilizar:

- Las pruebas más sensibles se utilizan para tamizaje (*screening*), al inicio del proceso diagnóstico, para descartar enfermedad, o cuando la oportunidad perdida de tratamiento presupone un gran riesgo.
- Las pruebas más específicas se utilizan para confirmación, al final del proceso diagnóstico, y para enfermedades cuyo tratamiento innecesario puede ser peligroso.

Cuando el resultado del estudio diagnóstico se expresa en una escala numérica continua, es necesario decidir a partir de qué valor de la prueba se considerará al sujeto como enfermo. Aquí ocurre algo interesante: modificar el punto de corte altera la capacidad operativa de la prueba:

- Si corremos el punto de corte hacia valores menos patológicos, ganamos en sensibilidad (incluimos más casos como enfermos), pero perdemos en especificidad (muchos de estos casos pueden ser falsos positivos).
- Si lo movemos hacia valores más patológicos, ganamos en especificidad (tendremos menos falsos positivos), pero perderemos sensibilidad (es posible que algunos enfermos escapen al diagnóstico).

a. Docencia e Investigación. Hospital Nacional de Pediatría "Prof. Dr. Juan P. Garrahan."

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

Conflicto de intereses:
Ninguno que declarar.

Recibido: 4-7-10
Aceptado: 28-7-10

Para entender esto con más claridad volvamos a nuestro artículo y observemos en la *Tabla 1* (Cuadro 2 en la publicación original[3]) cómo se modifica la capacidad operativa (medida por la razón de probabilidad o "*likelihood ratio*" que relaciona los verdaderos y falsos positivos y negativos) utilizando distintos puntos de corte de $PcapO_2$. (Como repaso de los conceptos del artículo anterior se pueden reconstruir con estos datos las tablas de 2 x 2 y calcular las medidas operativas para cada punto de corte).

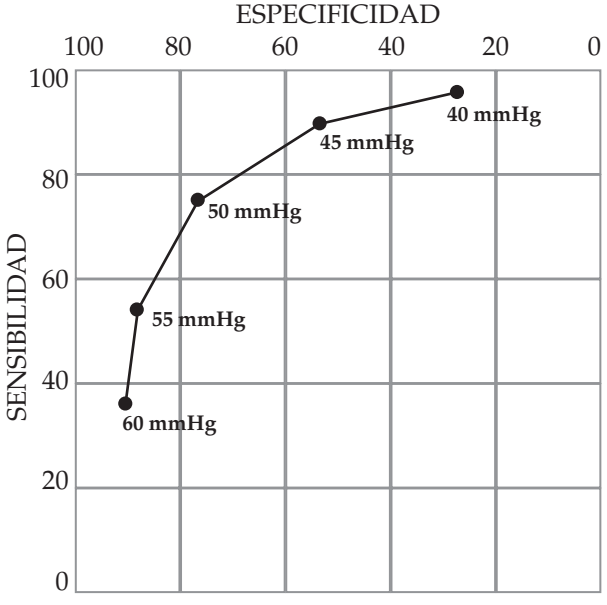
La elección del valor límite (punto de corte) óptimo de una prueba con resultado numérico implica balancear estas dos alternativas para maximizar la capacidad operativa del método (máxima sensibilidad y especificidad) y minimizar los errores "negociando" entre verdaderos positivos (sensibilidad) y falsos positivos (1 – especificidad). Para este proceso se utiliza un gráfico denominado curva ROC (*Receiver Operating Characteristic*) o curva de respuesta relativa.[4] Estas curvas son diagramas de correlación donde se relacionan la sensibilidad en las ordenadas contra la especificidad en las abscisas para los diferentes puntos de corte de un método. La principal utilidad de la curva es asistir en la toma de decisión acerca de cuál es el valor límite óptimo, que es aquél que más se acerca al extremo superior izquierdo del gráfico: máxima sensibilidad y especificidad, o máxima tasa de verdaderos positivos (TVP) con un mínimo de falsos positivos (TFP).

En la *Figura 1* (reproducida de nuestro artículo de gases capilares) se presenta esta curva ROC para los distintos puntos de corte de PO_2 capilar

TABLA 1. *Rendimiento diagnóstico del gas capilar contra el arterial para distintos valores límite (puntos de corte)*[3]

	Probabilidad después de la prueba - razón de probabilidad			
PcO_2	**Hiperoxemia**	**Normal**	**Razón de probabilidad**	**Probabilidad posprueba**
> 60	18= 0,375	3= 0,057	6,6:1	87%
> 55	27= 0,56	5= 0,096	5,8:1	85%
> 50	37= 0,77	11= 0,21	3,6:1	78%
> 45	44= 0,92	24= 0,46	2,0:1	67%
> 40	47= 0,98	39= 0,75	1,3:1	56%
Total	48	52		

FIGURA 1. *Curva ROC para distintos valores límite (puntos de corte) del gas capilar contra el arterial*[3]



para el diagnóstico de hiperoxemia. Observe que:

- en el eje Y se representa la sensibilidad (TVP) de manera creciente de 0 a 100%;
- en el eje X se grafica la especificidad (1 - TFP) en forma decreciente de 100 a 0% (también podría mostrarse en escala creciente de 0 a 100% pero como 1 – especificidad);
- a medida que aumenta la sensibilidad, la especificidad disminuye y viceversa;
- el valor de $PcapO_2$ de 50 mmHg es el punto que tiene mayor TVP y menor TFP (el que se ubica más cercano al ángulo superior izquierdo de la curva).

Cuando se elabora una curva ROC por medio de un programa informático estadístico, éste informa un valor para el área bajo la curva. El área total del gráfico es igual a 1; los valores posibles del área comprendida bajo la curva ROC oscilan entre 0 y 1. Esta medida es proporcional a la capacidad diagnóstica, de tal manera que cuanto mejor sea el desempeño de la prueba el área será mayor (más cercana a 1). Estas curvas y áreas bajo la curva permiten no sólo comparar puntos de corte de un método determinado sino también comparar el desempeño de dos o más métodos diagnósticos diferentes.

Como cierre, es importante destacar que este punto de corte es óptimo para el uso amplio de la prueba en todo el espectro de aplicaciones. Sin embargo, es perfectamente lícito utilizar un punto de corte diferente si se apunta a un uso particular de la prueba diagnóstica que requiera optimizar la sensibilidad (rastreo o *screening*), o bien maximizar la especificidad (confirmación diagnóstica). También es posible calcular el desempeño diagnóstico para distintas categorías o intervalos de valores de la prueba en lugar de utilizar un único punto de corte.[5]

Reglas de predicción clínica

El proceso diagnóstico es complejo y requiere una justa combinación de experiencia clínica y exámenes complementarios. El avance tecnológico ha permitido el desarrollo de innumerables prueba diagnósticas que van desde las muy simples hasta las sumamente sofisticadas; pero el campo de los instrumentos diagnósticos clínicos está aún relativamente virgen. Interpretar los resultados de las pruebas diagnósticas suele parecer bastante sencillo: basta conocer los valores normales de laboratorio o adquirir la habilidad visual requerida para los estudios de imágenes, sin olvidar el aspecto que analizamos en nuestro artículo previo respecto de la capacidad operativa de la prueba (su sensibilidad, especificidad y valores predictivos). En cambio, los signos y síntomas clínicos requieren otro enfoque. Primeramente, es necesario tener la destreza para detectarlos, ya sea mediante un examen físico apropiado o una anamnesis dirigida pero exhaustiva (habilidad que luchamos por conseguir durante nuestros primeros años de formación de postgrado…). Luego, corresponde ponderar su impacto sobre nuestros diagnósticos presuntivos, ya que las formas de presentación son sumamente variables y todos los signos clínicos no "pesan" lo mismo. Este último proceso suele ser el más "artesanal" de la medicina; tanto es así, que se va perfeccionando a lo largo de los años, es difícil de explicitar de manera precisa para poder transmitirlo a los más jóvenes, y se lo acaba expresando con el término "ojo clínico", que denota nuestra total incapacidad para estandarizarlo o sistematizarlo.

Por suerte, la Estadística puede brindar un aporte valioso en este campo. Así como el desarrollo tecnológico mejoró la calidad de las pruebas diagnósticas disponibles, haciéndolas más fáciles de aplicar, menos invasivas y menos subjetivas, el crecimiento de los métodos estadísticos multivariados gracias a la informática permite la construcción de instrumentos estandarizados y confiables de diagnóstico o pronóstico, procedimiento que está al alcance de cualquier investigador clínico que disponga del número suficiente de pacientes y asesoramiento estadístico. Estas herramientas, denominadas "reglas de predicción clínica" ("*clinical prediction rules*", "*decision rules*"), proveen un abordaje estructurado para diagnosticar una enfermedad o estimar el riesgo de un evento, y si se elaboran con la metodología adecuada, tienen la doble capacidad de ponderar la contribución individual de cada signo y poder ser aplicadas de manera eficaz para la toma de decisiones diagnóstica o terapéutica por profesionales con menor experiencia.

Metodología para la construcción de reglas de predicción clínica

Cuando utilizamos una prueba diagnóstica lo hacemos porque tenemos evidencia previa de que los pacientes con prueba positiva o con determinados valores tienen mayor probabilidad de estar enfermos; esta evidencia proviene de los estudios de validación de pruebas diagnósticas como las que analizamos en nuestro artículo anterior. Lo mismo ocurre con los síntomas y signos clínicos que, como ya vimos, también pueden ser considerados pruebas diagnósticas. Ahora bien: todos los

signos, síntomas o estudios complementarios no repercuten del mismo modo sobre nuestro diagnóstico; la integración de toda esta información (a menudo contradictoria, excepto en casos floridos) suele ser asistemática, basada en nuestra experiencia previa y muy artesanal. Sin embargo, podemos aplicar conceptos ya analizados en esta serie para mostrar cómo el desarrollo y uso de reglas de predicción clínica puede sistematizar nuestro proceso diagnóstico o pronóstico, haciéndolo más explícito y eficiente.

Para entrar en este tema es interesante describir las distintas etapas en la investigación del valor diagnóstico de un signo clínico o un examen complementario.[6] De manera similar a lo que ocurre en el estudio de nuevos fármacos, el desarrollo de pruebas diagnósticas atraviesa cuatro fases sucesivas:[7]

- *Fase I:* El primer paso es explorar si el resultado de la prueba es diferente entre pacientes con enfermedad conocida y sujetos sanos. Esto es básico, ya que si la prueba falla en distinguir sanos de enfermos, la investigación se detiene ahí.
- *Fase II:* El segundo paso es averiguar si los pacientes con determinados resultados de la prueba tienen mayor probabilidad de estar enfermos. En esta etapa también se utilizan sujetos de los extremos del espectro de enfermedad (sanos y enfermos conocidos), para aumentar la evidencia que avale que la prueba es promisoria.
- *Fase III:* La tercera etapa intenta establecer si la prueba distingue entre enfermos y sanos entre sujetos sospechados de tener la enfermedad. Aquí lo que se pretende es ver si la prueba es útil en la situación clínica real en la que se aplica para hacer diagnóstico en la gama "gris" del espectro de enfermedad, no ya para distinguir entre pacientes claramente enfermos o sanos. Esta es la fase de los estudios de validación que analizamos en nuestro artículo anterior, donde los resultados de la prueba se comparan contra los del estándar de referencia (el "*gold standard*") en un grupo de individuos de un espectro apropiado de la enfermedad.
- *Fase IV:* Finalmente, para decidir si vale la pena aplicar la prueba diagnóstica se necesita demostrar que su uso tiene algún impacto sobre la evolución de la enfermedad. Aquí, la prueba se analiza como una intervención diagnóstica, por lo que el diseño óptimo para verificar su eficacia es un ensayo clínico controlado y aleatorizado en el que se analice si el grupo al que se aplicó la prueba tiene mejores resultados de salud (es decir: si el diagnóstico más temprano o más eficaz contribuyó a mejorar el pronóstico de la enfermedad).

Ahora supongamos que queremos desarrollar un instrumento de diagnóstico o pronóstico utilizando un conjunto de signos clínicos y estudios complementarios, o sea una regla de predicción clínica. Nuestra línea de investigación incluiría básicamente 2 etapas:[8-11]

1. *Etapa de derivación*

La primera tarea sería seleccionar de una lista exhaustiva de potenciales predictores (extraídos de la bibliografía y la propia experiencia) aquellos asociados con mayor probabilidad de un determinado diagnóstico o pronóstico de la enfermedad: esta es la etapa de derivación de nuestra regla (equivalente a las fases I y II). Ya hemos visto antes cómo se estudian los factores de riesgo: utilizando idealmente una cohorte de sujetos (la denominada "muestra de derivación" o "*training set*"), podríamos identificar aquellos signos o factores asociados a la enfermedad; esto significa que sus medidas de efecto en el análisis bivariado, riesgo relativo (RR) u *odds ratio* (OR), son mayores de 1 con intervalos de confianza (IC) que no contienen el 1. Para controlar posibles sesgos de confusión o interacciones entre ellos, es conveniente incluir aquellos con diferencias significativas o límite ($p< 0,10$) en un análisis multivariado; este análisis nos proveerá la magnitud del efecto ajustada por todos los confundidores (o "covariables") incluidos en el modelo, expresada por el OR o RR "ajustados" que constituyen una medida del "peso independiente" de cada factor para el diagnóstico o pronóstico de interés. Con esta información, estamos en condiciones de armar nuestra regla de predicción o "*score*", otorgando a cada factor seleccionado como significativo un puntaje que sea proporcional a su medida de efecto ajustada (su peso independiente).

2. *Etapa de validación*

Una vez creado el "*score*" (y antes de utilizarlo para la toma de decisiones) debemos corroborar que, de verdad, mide lo que queremos que mida, es decir "validarlo". Esta es la etapa de validación (similar a la fase III de estudios de validación de métodos diagnósticos analizada en nuestro número anterior). Aquí el objetivo es verificar la capacidad operativa de la prueba (el puntaje o *score* creado) frente a algún estándar de referencia que nos proporcione la mejor certeza diagnóstica posible. Este "*gold standard*" puede ser un solo método diagnóstico o una combinación o secuen-

cia de pruebas valoradas en conjunto como una única prueba. Para el diseño de esta etapa caben todas las consideraciones de validez interna ya comentadas en nuestro artículo anterior: utilizar un diseño transversal, independiente y en lo posible con enmascarmiento ("ciego"), y un espectro de pacientes amplio y parecido al del futuro ámbito de aplicación del *score*. Un aspecto insoslayable de estos estudios es que deben realizarse sobre una nueva muestra de sujetos (la "muestra de validación" o "*testing set*"), que sustente la validez externa del instrumento (que es aplicable y eficaz en otros subconjuntos de sujetos similares a aquellos de los que se derivó). En esta etapa se evalúan la calibración del instrumento (la concordancia entre la probabilidad estimada del evento y la observada realmente) y su discriminación (la relación entre aciertos y errores evidenciable por el área bajo la curva ROC). Para la aplicación del puntaje, se puede elegir un único punto de corte (balanceado mediante una curva ROC o bien con máxima sensibilidad, ya que estos instrumentos se utilizan habitualmente como *screening*); alternativamente, se pueden establecer categorías de puntajes que representen alto o bajo riesgo del evento. Los resultados analizados en esta fase serán las medidas de capacidad operativa para cada punto de corte o cada categoría de puntajes: sensibilidad (S), especificidad (E), valores predictivos (VP) y razones de probabilidad (*likelihood ratios*, LR) positivos y negativos con sus respectivos IC 95%. Si nuestro puntaje predice o diagnostica bien, quedará así validado y, en líneas generales, podría ser utilizado en poblaciones similares.

Veamos un ejemplo:

En un artículo publicado en *Critical Care Medicine* en 1988,[12] Pollack y col. comunican la derivación y validación del conocido *Pediatric Risk of Mortality* (PRISM) que es una escala o puntaje para pronosticar riesgo de muerte en terapia intensiva pediátrica. Hasta ese momento, el riesgo de muerte en este tipo de pacientes se estimaba con el *Physiologic Stability Index* (PSI) elaborado mediante un consenso de expertos, que valoraba 34 variables fisiológicas. Para intentar reducir este gran número de factores requeridos para estimar el pronóstico, los autores estudiaron una cohorte multicéntrica que abarcó nueve unidades de cuidados intensivos (UCI) pediátricas, utilizando la mitad de los datos para la derivación ("*estimation set*") y la otra mitad para la validación ("*validation set*"). Para la creación del PRISM *score* se utilizó un método multivariado (regresión logística) que seleccionó las 14 variables que componen el instrumento. Éste fue sometido a un proceso de validación que mostró una predicción muy similar a la del PSI (observable en la curva ROC, con un área bajo la curva de 0,92).

Utilidad de las reglas de predicción clínica

Los puntajes clínicos (*scores*) bien diseñados y validados tienen un gran valor en el proceso diagnóstico o la estimación del pronóstico, facilitando la elección de un determinado manejo terapéutico o la información al paciente sobre el curso de su enfermedad. Constituyen así herramientas valiosas para apoyar la toma de decisiones diagnósticas o terapéuticas por profesionales de diverso grado de experiencia, ya que habitualmente requieren la medición de unas pocas variables sencillas de obtener (a veces marcadores o subrogantes de otras más difíciles de medir, como el color de la piel en el puntaje de Apgar).

Por otra parte, su revalidación en distintos ámbitos amplía su validez externa, sustentando su aplicabilidad a poblaciones similares, aunque algo diferentes de aquella en la cual se crearon. (El puntaje PRISM, por ejemplo, fue revalidado en una UCI pediátrica en India,[13] mostrando un menor desempeño (área bajo la curva ROC de 0,80) atribuible a diferencias poblacionales y de recursos).

Las escalas cuantitativas adecuadamente validadas sirven, además, para estandarizar la clasificación de enfermedades, categorizar con precisión grupos de riesgo o gravedad o mostrar variaciones evolutivas (como los puntajes de APACHE o de Glasgow). Todas estas utilidades tienen aplicación no sólo en el ámbito asistencial, sino también en el campo de la investigación clínica.

El uso de reglas de decisión puede influir sobre el manejo clínico mejorando los resultados de salud de los pacientes, la calidad de la atención o su costo-efectividad. La valoración de este impacto implica considerar el uso de la regla como una intervención (fase IV mencionada anteriormente) y evaluar su eficacia mediante diseños experimentales (ensayos aleatorizados) o cuasi-experimentales (antes-después).[14]

La estimación de riesgo por medio de reglas de predicción repercute también sobre la organización y gestión de servicios de salud; puede utilizarse para la asignación racional de recursos y permite la evaluación comparativa de la calidad de atención entre distintos centros o diferentes períodos de un mismo centro (como el *Clinical Risk Index for Babies* o *CRIB score*, útil para comparar

mortalidad y *performance* entre unidades de cuidados intensivos neonatales).

Finalmente, para que estos beneficios puedan observarse, es necesario que la regla sea simple, objetiva, válida y útil para que sea incorporada a la práctica. Conocer las bases metodológicas de su desarrollo y revalidarlas en el ámbito de la propia tarea pueden promover la utilización de reglas de predicción por parte de los profesionales de la salud.

Los instrumentos de diagnóstico adecuadamente diseñados y validados son escasos en la bibliografía pediátrica. La disponibilidad de métodos estadísticos para la selección y ajuste de variables hacen del desarrollo de reglas de predicción un campo interesante de investigación, que espero haber promovido entre los pediatras con este artículo. ■

BIBLIOGRAFÍA

1. Demirdjian G, Berlín V, Rowensztein H. Pediatría basada en la evidencia. Estudios de diagnóstico (1ª Parte). *Arch Argent Pediatr* 2009;107(6):527-535.
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. Hinojosa-Pérez JO, Treviño Báez JD. Utilidad de la gasometría capilar para detector hiperoxemia en el recién nacido grave. *Bol Med Hosp Infant Mex* 1999;56(2):93-96.
4. Altman DG, Bland LM. Diagnostic tests 3: receiver operating characteristic plot. *BMJ* 1994;309:188.
5. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, et al. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002;324:669-671.
6. Ferrero F. Reglas de predicción clínica. *Arch Argent Pediatr* 2010;108(1):6-7.
7. Sackett DL, Haynes RB. Evidence base of clinical diagnosis. The architecture of diagnostic research. *BMJ* 2002;324: 539-541.
8. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: application and methodological standards. *N Engl J Med* 1985;313:793-799.
9. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;277:488-494.
10. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, et al. Prognosis and prognostic research: what, why and how? *BMJ* 2009;339:b375.
11. Wade A. Derivation versus validation. *Arch Dis Child* 2000; 83:459-460.
12. Pollack MM, Ruttimann UE, Getson PR. Pediatric Risk of Mortality (PRISM) Score. *Crit Care Med* 1988;16:1110-1116.
13. Thukral A, Lodha R, Irshad M, Arora NK. Performance of Pediatric Risk of Mortality (PRISM), Pediatric Index of Mortality (PIM), and PIM2 in a pediatric intensive care unit in a developing country. *Pediatr Crit Care Med* 2006;7:356-361.
14. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201-209.

"Los hombres que son capaces de sacrificar libertad en aras de la seguridad
no merecen ninguna de las dos."

Benjamín Franklin