Rev. Chil. Pediatr. 66 (6); 335-340, 1995

Pruebas diagnósticas: algunos aspectos cuantitativos para su desarrollo e interpretación

Patricio Herrera L.¹; Gastón Duffau T.¹

Sensitivity and specificity of diagnostic tests

Diagnostic tests are common and useful tools for clinical use since they give way to diagnostic, therapeutic and prognostic decisions concerning individual patients. Most guidelines for test development and application usually deal with qualitative more than quantitative requirements. This study is addressed to examine some quantitative issues concerning mainly the estimation of expected sensitivity (and specificity) of fixed indexes on which predictive value calculations rely. Sample size required for accurate and precise estimations of sensitivity or specificity is shown to be determinant of potential bias source, particularly when tests application results lead to critical decisions. Definitions of "prevalence" according to whether it is regarded as the proportion of cases in the test development, the whole population or a particular population sample are also reviewed.

[Key words: diagnosis, diagnostic tests.]

Las pruebas diagnósticas constituyen una materia en permanente evolución, son objeto frecuente de publicaciones y tema atractivo para investigadores clínicos porque traducen, en gran medida, el avance científico en términos de progreso tecnológico aplicable a la atención de enfermos.

La utilidad de las pruebas diagnósticas reside en que, dadas las características intrínsecas propias de cada una de ellas (la constancia de los valores de sus índices fijos), y atendidos ciertos requisitos al ensayarla o reproducirla, proporcionan información —expresada en probabilidades— que permite al médico tomar decisiones más o menos graves en aspectos clínicos (diagnósticos, pronósticos o terapéuticos) ante determinado paciente. Las condiciones para evaluar una prueba antes de aplicar sus resultados al uso clínico, han sido revisadas¹⁻⁶ y están dirigidas a evitar sesgos en la medición de sus características propias y en la interpretación de sus resultados.

Aunque los resultados de estas pruebas son numéricos (generalmente probabilidades expresadas como porcentajes), la mayoría de los au-

tores han centrado la atención sobre los aspectos más bien cualitativos que podrían distorsionarlos, que han sido resumidos en general, y de modo más bien básico, por Sackett *et al.*¹

El análisis de diversas publicaciones, que proponen el uso de pruebas o procedimientos con fines diagnósticos, permite identificar fuentes de sesgo importantes en el ensayo e interpretación de estas pruebas. Excepto comentarios editoriales aislados⁷, no parece haber en la literatura clínica una reacción de expertos en términos de la presentación de un cuerpo de información sistematizada, como los citados más arriba, sobre requisitos, además de los citados, para el ensayo e interpretación de pruebas diagnósticas.

El propósito de la presente comunicación es analizar algunos aspectos técnicos cualitativos y cuantitativos relevantes en el ensayo de estas pruebas, para contribuir —en nuestro medio— al conocimiento de esta materia⁸⁻¹¹. Aunque por razones circunstanciales la literatura estudiada se refiere al campo de la pediatría y, dentro de éste, a la infectología, los aspectos abordados tienen utilidad general. La revisión bibliográfica no fue exhaustiva porque el propósito no era hacer una evaluación global de la literatura sobre el tema sino ofrecer un análisis ilustrativo

1. Universidad de Chile, Facultad de Medicina, Departamento de Pediatría, Campus Norte.

sobre el mismo. Los ejemplos citados también tienen sólo este propósito.

Para simplificar la identificación de los elementos mencionados en esta comunicación, se presenta el esquema común de resultados e interpretación de pruebas diagnósticas (tabla 1) y los requisitos básicos para su evaluación metodológica (tabla 2).

Las fuentes de sesgo en la medición e interpretación del comportamiento de las pruebas diagnósticas se relacionan con aspectos observables, ya desde el punto de vista intuitivo, en la construcción del diseño para evaluar dicho comportamiento. Esta comunicación se refiere al tamaño muestral, en particular respecto al número de pacientes en la categoría de "casos" o "verdaderos positivos".

Para evitar confusiones, debemos puntualizar algunos conceptos y términos básicos respecto a los cuales se requiere consenso, aun cuando el lector familiarizado con el tema podría oponer cierto grado de controversia.

El primero es el de los índices fijos (sensibilidad y especificidad) de determinada prueba (en particular, se hace referencia a las pruebas cuyos resultados pueden ser sólo "positivo" y "negativo", incluyendo una escala de valores en que se elige un "punto de corte", que convierte una serie, de una variable continua, en dos categorías: la de valores considerados "positivos" y la de los valores "negativos"). Para hacer predicciones (calcular los valores predictivos o índices variables) es indispensable que los índices fijos sean constantes a través de las aplicaciones de dicha prueba. Los requisitos más generales, resumidos y comentados por Sackett¹, están dirigidos, precisamente, a asegurar esta "inmutabilidad" de los valores de sensibilidad y especificidad. Esto conlleva que la prueba tiene valores propios de índices fijos (idealmente), independientes del nivel y población en que se aplica, lo que permite deducir o calcular los índices variables o "predictivos", cuando la prueba es aplicada a individuos de una población en que la prevalencia de la enfermedad para la que está diseñada es diferente de la que aparece en la publicación donde se ensayó la prueba. Los valores predictivos entregados por los autores del ensayo de una prueba deberán ajustarse, al aplicarla a individuos de una población, según prevalencia de la enfermedad en esa población. En principio, no hay procedimiento alguno, si la

Tabla 1

Interpretación de los resultados
de una prueba diagnóstica

Resultado de la prueba	Condición de la enfermedad	
	Presente	Ausente
Positivo	a	b
Negativo	c	d
Total	a + c	b + d
a: verdaderos positivos; b: falsos positivos		
c: falsos negativos; d: verdaderos negativos		
Indices fijos: sensibilidad: $a/a + c$; especificidad: $d/b + d$		
Indices variables: valor predictivo positivo: $a/a + b$		
valor predictivo negativo: $d/c + d$		
prevalencia: $a + c/a + b + c + d$		

Tabla 2

Guía para decidir sobre la validez de una nueva
prueba diagnóstica*

1. ¿Hubo una comparación independiente, ciega, con un patrón diagnóstico de referencia?
2. ¿Incluyó la muestra de pacientes un espectro adecuado de los enfermos y de los individuos con enfermedades diferentes pero que se pueden confundir con los primeros?
3. ¿Se describió adecuadamente el lugar y filtro de referencia de los pacientes en que se hizo el estudio?
4. ¿Se determinó la precisión y reproducibilidad de la prueba?
5. ¿Se definió aceptablemente el término "normal"?
6. Si la prueba se propuso como parte de un conjunto o secuencia ¿se determinó su contribución a la validez del conjunto o secuencia?
7. ¿Se describió la táctica para realizar la prueba en detalle suficiente para poder reproducirla exactamente?
8. ¿Se determinó la utilidad o aplicabilidad real de la prueba?

* Adaptado de Sackett DL, Haynes RB, Tugwell P. Clinical Epidemiology, Little Brown and Co. Boston 1985: 293.

prueba es reproducida adecuadamente, que se pueda agregar a ésta para cambiar el valor de sus índices fijos, ya que de haberlo se convertiría en un instrumento no confiable.

Antes de considerar cualquier cálculo de los índices variables, cabe examinar el significado de los índices fijos, los cuales son, separadamente, una tasa o proporción. Así, la sensibilidad, representada por $a/(a + c)$ en la tabla 1, es

la proporción en que la prueba identifica los verdaderos positivos de una población de individuos en que el patrón de referencia o "estándar ideal" permitirá clasificar como el total de los "casos". En este sentido, y para los efectos de este análisis, debe considerarse la sensibilidad como una proporción independiente de la representada por la especificidad (aunque en el comportamiento de estas pruebas sabemos que existe una relación intrínseca entre ambas proporciones). En otras palabras, y a modo de ejemplo, si una prueba detecta 80% de los casos de enfermedad, entre 100 verdaderos portadores de ésta, podemos aceptar que la sensibilidad de la prueba es 80%.

Intuitivamente, nos resultará fácil reconocer que el valor de la estimación de la sensibilidad depende del número de casos en que fue aplicada: 80/100 parece (y es) más confiable que 4/5 (80%). Surge de inmediato la pregunta ¿cuántos individuos que, efectivamente, padecen la enfermedad en estudio se necesitarán para que la proporción "sensibilidad" sea confiable (v gr 80% y no 60, 85 u otra cifra)? La respuesta debe buscarse en las propiedades de la distribución binomial en casos en que la proporción es menor de 0,30, en base a la cual se pueden construir límites o intervalos de confianza¹². Para el ejemplo dado, estos límites —con 95% de seguridad— son aproximadamente 72,2 y 87,8%. Es decir, con un margen de error de 5%, la sensibilidad puede ser cualquier valor entre estos dos extremos. Según el propósito con que se emplee la prueba diagnóstica —como se comenta más adelante—, puede ser importante ponerse en el caso, por ejemplo, que la sensibilidad real de la prueba sea sólo 72,2%. En términos epidemiológicos, ello implica incapacidad de ésta para reconocer 27,8% de los casos y no 20% como señala el estimador puntual inicial (80%).

Dada esta dificultad, que afectará especialmente la cifra o proporción llamada "valor predictivo negativo", como consecuencias extremas cuando la sensibilidad es estimada como cercana a 100% (tabla 3), cabría intentar resolver la cuestión del tamaño muestral (para la estimación de la sensibilidad), buscando estimar con mayor exactitud la cifra real de la sensibilidad intrínseca de la prueba y, a la vez, hacer más estrecho el intervalo de confianza, lo que equivale a identificar un valor de sensibilidad más preciso con intervalo de confianza más es-

Tabla 3

Ejemplo de efecto sobre los valores predictivos de la estimación de la sensibilidad con tamaño muestral restringido

A: Condición de la enfermedad			
Resultado	Presente	Ausente	Total
Positivo	31	7	38
Negativo	0	23	23
Total	31	30	61

Sensibilidad 100% (86,2 a 99,7, IC 95%); especificidad 76,7%; VPP 81,6%; VPN 100 %

B: Condición de la enfermedad			
Resultado	Presente	Ausente	Total
Positivo	27	7	34
Negativo	4	23	27
Total	31	30	61

Sensibilidad 87,1%; especificidad 76,6%; VPP 76,6%; VPN 79,4%

C: Condición de la enfermedad			
Resultado	Presente	Ausente	Total
Positivo	27	70	97
Negativo	4	230	234
Total	31	300	331

Sensibilidad 87,1; especificidad 76,7%; VPP 27,8%; VPN 98,2%

VPP: valor predictivo positivo; VPN: valor predictivo negativo.

Notas: en B se ajustaron los valores de los índices fijos y se calcularon los variables según un valor aproximado —para evitar "decimales de pacientes"— al del límite inferior del intervalo de confianza de 95%, previendo que la sensibilidad real de la prueba no sea 100%; los casos no detectados en este caso son 4. En C, pese a mejorar el valor predictivo negativo, por aumento de los "controles", siguen quedando 4 casos sin detectar.

trecho (exactitud o validez de una medición es el grado en que los resultados de una medición corresponden al estado verdadero del fenómeno que se está midiendo; precisión o confiabilidad es el grado en que mediciones repetidas de un fenómeno relativamente estable caen cerca unas de otras¹³).

¿Qué implicaciones tienen estas consideraciones?: una decisión imprescindible al estudiar el comportamiento de una prueba diagnóstica se

refiere a cuál es el problema que esperamos que ella nos resuelva. La mayoría de las pruebas se usan para detectar (o confirmar, que es otra categoría de reconocer los verdaderos positivos) o descartar la presencia de verdaderos positivos. Algunas (la minoría) sirven eficientemente para ambos objetivos. Otra consideración importante es examinar la situación en la que se usará la prueba diagnóstica, particularmente respecto al grado de gravedad de la decisión a que den lugar sus resultados. En la situación más demostrativa, cual es la de decisiones críticas (tratar o no un caso que, sin tratamiento, muy probablemente evolucionará hacia la muerte o agravación), cuando la sensibilidad es 100%, el valor predictivo negativo será 100% (independientemente de la especificidad). En otras palabras, ante un caso dado, si el resultado de la prueba es negativo, ello implica que la probabilidad de estar ante un caso verdadero de la enfermedad-problema es igual a 0, es decir, si la enfermedad requiere tratamiento, este resultado implica no tratarla. Si de todos modos, porque la condición es muy grave, el tratamiento tiene muy bajo coste, etc., se decide tratar al paciente, a pesar del resultado de la prueba (en condiciones ideales, lo que incluye 50% de probabilidad de acertar o no), ésta no sirve desde el momento en que estamos tomando decisiones en base a otros antecedentes que nos sugieren ignorar los resultados en cuestión. La situación inversa se plantea cuando la especificidad es 100%. Si la prueba es positiva, la probabilidad de que el enfermo con este resultado no tenga la enfermedad es igual a 0. En este caso, la situación extrema subraya la importancia de este hecho, si la decisión de tratar al enfermo tiene consecuencias graves por sus riesgos, costos, u otras razones.

Sobre todo para estos casos extremos, cuando los valores predictivos son matemáticamente determinantes en situaciones críticas, los índices fijos deben ser lo más exactos y precisos posible. La única manera de lograrlo es calculando un tamaño muestral que garantice razonablemente los dos objetivos relacionados: exactitud y precisión. El cálculo del tamaño muestral para determinar un índice fijo requiere alguna idea sobre su valor (p.e. por medio de estudios-piloto o estudios previos llevados a cabo en muestras relativamente pequeñas). No debe esperarse que el valor de la nueva estimación, hecha con una muestra "adecuada", sea igual al de

la prueba piloto sino diferente (mayor o menor). Si resulta menor, parece obvio que debe considerarse no utilizar la prueba. El riesgo de los tamaños muestrales insuficientes está en la determinación de distorsiones en los valores predictivos calculados, en base a índices fijos inexactos, con consecuencias eventualmente graves para el paciente.

Otro aspecto de alguna manera relacionado con el anterior deriva de una confusión provocada por el término "prevalencia" (de la enfermedad) en el lenguaje sobre pruebas diagnósticas. En realidad hay dos "prevalencias": la de la enfermedad en estudio en la población en que se ensaya la prueba ($a+c/a+b+c+d$), que es —en cierto modo— arbitraria o artificial, porque generalmente es muy elevada y construida por el investigador. La otra "prevalencia" importante es la tasa de la enfermedad en la población general a la que se aplicará la prueba una vez estudiada, que generalmente es más baja que la expuesta por los autores que proponen la prueba diagnóstica. Esta segunda prevalencia, natural o poblacional, es la que se debe tener en cuenta al decidir adoptar una prueba para aplicarla a una población, previendo necesidades de ajustar los valores predictivos proporcionados por los autores del estudio de la prueba, para derivar los valores predictivos en individuos de la población no experimental (p.e. si la prueba pretende detectar infección urinaria, los valores predictivos resultantes —a partir de la sensibilidad y especificidad de dicha prueba— dependen de la proporción de casos de infección urinaria que hay en la población que acude a un consultorio no especializado).

En el trabajo clínico hay una tercera prevalencia que es la probabilidad de la enfermedad en estudio en el paciente individual al que se practicará la prueba para tomar una decisión¹. Esta prevalencia es eventualmente diferente de la poblacional, si el paciente pertenece a un subgrupo de ella (en el ejemplo anterior es la probabilidad específica de padecer infección urinaria que tiene un individuo en particular —que consulta por un síndrome que obliga a descartar esta enfermedad— por el hecho de pertenecer a un subgrupo de la población determinado, por ejemplo, por sexo o edad, determinantes de tasas de la enfermedad diferentes respecto a la de la población general). Esta prevalencia y la anterior también deben ser estimadas en forma

exacta al ser aplicada la prueba a un individuo en particular para hacer una predicción en base al resultado de la prueba en él.

Ninguna prevalencia afectará directamente los valores de los índices fijos de una prueba, sino sólo de los variables o predictivos. Sin embargo, como ocurre con frecuencia, los verdaderos enfermos (casos) son relativamente pocos, mientras los controles suelen ser más abundantes, lo que tiende a reducir el número de los verdaderos enfermos a uno más bien pequeño, lo cual es compensado por un número más grande de controles. Esta exigencia tiende, inadvertida e indirectamente, a estrechar el espectro clínico de uno o ambos grupos y a reducir el número total de casos verdaderos de la enfermedad, afectando la exactitud y precisión de la estimación de los índices fijos. Ambos fenómenos distorsionarán el comportamiento e interpretación de la prueba hasta un punto que impedirá confiar en sus resultados y tomar decisiones clínicas, sobre todo cuando las circunstancias las convierten en graves.

Es oportuno destacar que el ensayo de pruebas diagnósticas, para determinar sus índices fijos, se lleva a cabo en poblaciones. Su utilidad clínica implica su aplicación a personas en particular. Por ello, en el ensayo de cualquier prueba es necesario buscar la mayor exactitud y precisión posibles ya que, sea cual sea el objetivo de su uso, el efecto final será una decisión clínica que afecta a personas. Sin embargo, es conveniente tener presente que no siempre se necesita que (y pocas pruebas lo cumplen) la sensibilidad o la especificidad de una prueba sean 100%. Hay situaciones en que un margen razonable de error de los valores predictivos (basados en índices fijos exactos), como es el caso en que éstos son inferiores a 100%, aún tiene utilidad, si los errores en la clasificación de algunos enfermos no implican peligro clínico, costos materiales o intangibles importantes, como es el caso de aquellas enfermedades que pueden ser objeto de seguimiento u observación directa para corregir las fallas de clasificación. Hay otros casos, en cambio, en que las decisiones deben tomarse por la inminencia de consecuencias graves. Es en esta última situación donde uno o ambos valores predictivos útiles deben ser iguales a 100%. En el primer caso, el número de individuos necesario para lograr un grado razonable de exactitud del o los índices fijos

esperados, en cada uno de los grupos ("casos" y "no casos" o "controles", respectivamente), es no menor de 350. Si, en cambio, el índice fijo esperado es 100%, este número es mayor de 500¹².

Los ensayos con muestras insuficientes de "casos" o "controles" (v gr, para estimar la sensibilidad o especificidad) son aportes importantes pero no definitivos para resolver el dilema que, en último caso, implica tomar decisiones sobre individuos en particular, porque tiene el carácter de estudio preliminar o "piloto".

Resumen

Las pruebas de diagnóstico son útiles en la clínica pues facilitan la toma de decisiones en la identificación del problema, el tratamiento y la formulación del pronóstico, entre otras, en pacientes individuales. La mayoría de las recomendaciones para el desarrollo, evaluación y aplicación de ellas enfatizan más los aspectos cualitativos que los cuantitativos de sus requerimientos. Este estudio pretende examinar algunos asuntos cuantitativos relativos principalmente a la estimación de la sensibilidad y especificidad esperadas de los índices fijos sobre los cuales se basan los cálculos de los valores predictivos. El tamaño muestral necesario para estimar con exactitud y precisión la sensibilidad y la especificidad es determinante como fuente potencial de sesgos, particularmente cuando la aplicación de los resultados de una prueba conduce a decisiones críticas. Las definiciones de prevalencia también son objeto de confusión ya que pueden referirse a proporción de casos en el desarrollo de una prueba, a toda la población o sólo a una muestra de ella.

(Palabras clave: diagnóstico, pruebas de diagnóstico.)

Referencias

1. Sackett DL, Haynes RB, Tugwell P: Clinical Epidemiology: A basic science for clinical medicine. Little Brown & Co, Boston, 1985.
2. Ingelfinger JA, Mosteller FM, Thibodeau LA, Ware JH: Biostatistics in clinical medicine. MacMillan Publ. Co, NY 2ª Ed., 1987.
3. Weinstein MC, Fineberg HV: Clinical decision analysis. WB Saunders Co, 1980.

4. *Riegelman RK*: Studying a study and testing a test. Little Brown & Co Boston, 1981.
5. *Feinstein AR*: Clinical biostatistics. XXI: On the Sensitivity, specificity, and discrimination of diagnostic tests. *Clin Pharmacol Ther* 1975; 17: 104-116.
6. *Morton RF, Hebel JR*: A study guide to epidemiology and biostatistics. University Park Press, Baltimore 2ª Ed. 1984.
7. *McCarthy PL*: The febrile infant. *Pediatrics* 1994; 397-399.
8. *Baker MD, Bell LM, Avner JR*: Outpatient management without antibiotics of fever in selected infants. *NEJM* 1993; 329: 1437-41.
9. *Santolaya ME, Cofré J, Beresi V*: C-Reactive protein: A valuable aid for the management of febrile children with cancer and neutropenia. *Clin Infect Dis* 1994; 18: 589-95.
10. *Rodríguez G, Boehme C, Soto L, et al.*: Diagnóstico de meningitis bacteriana aguda por aglutinación con partículas de látex. *Rev Med Chile* 1993; 121: 41-45.
11. *Casanueva V, Cid X, Cavicchioli G, Oelker M, Cofré J*: Adenosina deaminasa en fiebre tifoidea y otras enfermedades febriles. *Rev Chil Pediatr* 1991; 62: 221-26.
12. *Fleiss JL*: Statistical methods for rates and proportions John Wiley & Sons NY 2ª Ed. 1981.
13. *Fletcher RH, Fletcher SW, Wagner EH*: Clinical Epidemiology. The essentials. Williams & Wilkins, Baltimore 2ª Ed. 1988.

AVISO A LOS AUTORES

Con el objeto de acelerar la edición de los trabajos se solicita encarecidamente a los autores que disponen de computador con procesadores de texto IBM compatibles, enviar –en lo posible junto con las dos copias impresas de su material–, otra en disco flexible (tamaño 3,5" o 5,25", de doble o alta densidad) del mismo texto. El disco será devuelto de inmediato por correo certificado o por mano al autor corresponsal.