Revista Colombiana de Obstetricia y Ginecología
ISSN: 0034-7434
rcog@fecolsog.org
Federación Colombiana de Asociaciones de Obstetricia y Ginecología
Colombia

Castro-Jiménez, Miguel Ángel; Cabrera-Rodríguez, Daladier; Castro-Jiménez, María Isabel
EVALUACIÓN DE TECNOLOGÍAS DIAGNÓSTICAS: CONCEPTOS BÁSICOS EN UN ESTUDIO CON MUESTREO TRANSVERSAL
Revista Colombiana de Obstetricia y Ginecología, vol. 58, núm. 1, 2007, pp. 45-52
Federación Colombiana de Asociaciones de Obstetricia y Ginecología
Bogotá, Colombia

Revista Colombiana de Obstetricia y Ginecología Vol. 58 No. 1 • 2007 • (45-52)

EDUCACIÓN MÉDICA

EVALUACIÓN DE TECNOLOGÍAS DIAGNÓSTICAS: CONCEPTOS BÁSICOS EN UN ESTUDIO CON MUESTREO TRANSVERSAL

Assessing a diagnostic test: basic concepts in a study with naturalistic sampling

Miguel Ángel Castro-Jiménez, M.D., MSc*, Daladier Cabrera-Rodríguez, M.D.**, María Isabel Castro-Jiménez***

Recibido: agosto 30/06 - Revisado: enero 23/07 - Aceptado: febrero 7/07

RESUMEN

Los estudios de evaluación de pruebas diagnósticas son un diseño científico que nos ayuda a determinar la validez y la reproducibilidad de los procedimientos que pueden ser usados en la práctica clínica por médicos y otros profesionales de salud. Este es un artículo de formación en epidemiología para estudiantes y profesionales de la salud y sus objetivos son: a) resumir los conceptos utilizados durante un estudio de evaluación de tecnologías diagnósticas y b) explicar cómo calcular e interpretar las medidas usadas durante la realización de un diseño con muestreo transversal, enfatizando en el significado de términos como sensibilidad, especificidad y valores predictivos de la prueba.

Palabras clave: evaluación de tecnología biomédica, sensibilidad y especificidad, valor predictivo de las pruebas, validez de las pruebas.

SUMMARY

Diagnostic test evaluation studies represent a scientific design helping to determine the validity and reliability of the procedures which can be used in clinical practice by doctors and other health professionals. This article deals with medical training in epidemiology. Its objectives are summarising concepts used when evaluating a new diagnostic test and explaining how to calculate and interpret the measurements used making a design with naturalistic sampling, emphasizing the meaning of test sensitivity, specificity and predictive value.

Key words: biomedical technology assessment, sensitivity and specificity, predictive value of test, validity of tests.

INTRODUCCIÓN

Los estudios dirigidos a evaluar las tecnologías diagnósticas médicas ayudan a definir la validez y la reproducibilidad de los procedimientos que son utilizados por el personal de la salud para realizar la mejor aproximación posible a la condición real de sus pacientes. El uso indiscriminado de estas pruebas produce riesgos injustificados para la vida de quienes

* Magíster en Epidemiología. Departamento de Salud Pública, Centro de Investigaciones Epidemiológicas, Facultad de Salud, Universidad Industrial de Santander. Bucaramanga, Colombia Correo electrónico: mcastro2505@yahoo.es; mcastro@ins.gov.co

** Coordinador Médico Salud Social IPS S.A. Telefax: (57) 7 647 6850. Bucaramanga, Colombia. Correo electrónico: dalasalud@hotmail.com

*** Fisioterapeuta. Bucaramanga, Colombia. Correo electrónico: micastro70@yahoo.es

han sido expuestos y, en forma adicional, un aumento excesivo de los costos de atención,[1] pudiendo llevar al colapso a cualquier sistema de salud. Para evitar estas situaciones cada prueba debe seguir un proceso de evaluación antes de ser usada ampliamente en la población, de una manera similar a lo que ocurre cuando se desea introducir un nuevo medicamento al mercado.

Los propósitos de esta revisión son: a) realizar un resumen de los conceptos básicos que deben tenerse en cuenta cuando se realiza un estudio de tecnologías diagnósticas y b) describir cómo se calculan e interpretan las medidas de mayor uso cuando se ha planteado un diseño con muestreo transversal.[2] Este documento se basa en las publicaciones metodológicas más importantes en el tema,[1-5] a las que también puede dirigirse el lector interesado en ampliar información acerca de los conceptos, justificación y análisis de este y otros tipos de muestreo.

Definición de algunos términos de uso común durante la realización de estudios de tecnologías diagnósticas en salud

Existen tres tipos de muestreo posibles durante la realización de un estudio de tecnologías diagnósticas y en los que, aunque la interpretación final de los resultados es similar, deben tenerse en cuenta las diferencias en las fórmulas del cálculo. El diseño que utiliza muestreo transversal es aquel tipo de estudio en el que tanto la prueba considerada patrón de oro como la que se encuentra en evaluación se aplica a todos los individuos y cuyos cálculos se realizan como se explica más adelante. Los estudios con un muestreo prospectivo son aquellos en los que a todos los individuos del estudio se les realiza la prueba que está en evaluación pero solo a una parte de ellos (positivos y negativos) se les aplica el patrón de oro; mientras que, en el diseño con muestro retrospectivo se les aplica a todos el patrón de oro y a una parte de ellos (positivos y negativos) se les realiza la prueba en evaluación. Algunos términos que deben ser tenidos en cuenta durante la realización de este tipo de estudio son:

- El desorden (en inglés, *disorder*): es un término utilizado para referirse a una enfermedad o disfunción y que es equivalente a la alteración de las funciones físicas o mentales normales; es una característica que el paciente presenta o no en un momento específico.
- El patrón de oro (en inglés, *gold standard*): es la prueba o conjunto de pruebas que se considera la mejor disponible a la fecha para realizar un diagnóstico determinado. Debido a que no existe infalibilidad en el patrón de oro, la prueba que hoy consideremos como la mejor puede dejar de serlo con el tiempo y ser reemplazada por otra que esté en evaluación.
- El período de la prueba (en inglés, *period of testing*): es el tiempo comprendido entre el momento de decidir a cuál examen se va a someter un paciente y aquel en el que se ha obtenido toda la información pertinente a sus resultados.
- El período de seguimiento (en inglés, *follow-up period*): es el tiempo que sigue al período de prueba en aquellos casos en los que las tecnologías empleadas se utilizan como predictoras, mas no diagnósticas, de una enfermedad.
- La prueba diagnóstica (en inglés, *diagnostic test*): es el procedimiento o examen que permite que el diagnóstico sea obtenido en el período de prueba. Por ejemplo, Singh[6] estudió las características de una prueba rápida para el diagnóstico de malaria, llamada ParaHIT f, en algunas poblaciones de India que tenían prevalencias variables de la enfermedad; luego de comparar con el patrón de oro, que para este estudio fue el examen microscópico de muestras de sangre, el autor informó sensibilidades y especificidades que variaban según el lugar, pero que superaron el 80% de manera global.
- La prueba pronóstica (en inglés, *prognostic test*): es el procedimiento o tecnología que permite que el diagnóstico sea obtenido en algún momento durante el período de seguimiento. Por ejemplo, Tierney et al.[7] evaluaron la eficacia de las pruebas neuropsicológicas para detectar a los individuos

que, siendo sanos al aplicar las pruebas, tenían un mayor riesgo de padecer demencia de Alzheimer durante los 10 años siguientes. Para una prueba determinada, los autores informaron una sensibilidad del 73% y una especificidad del 70% luego de 10 años de seguimiento. En este caso, con el resultado de la prueba no se hizo el diagnóstico de enfermedad de Alzheimer pero se logró identificar a los individuos que tenían un mayor riesgo de desarrollar esta patología luego de algún tiempo.

- Curva Receptor-Operador (en inglés, *Receiver Operating Characteristic Curve –ROC–*): es una representación gráfica de la sensibilidad contra el complemento de la especificidad (proporción de positivos falsos con respecto a todos los negativos al diagnóstico) para los posibles puntos de corte de una prueba diagnóstica.[8]
- Validez (en inglés, *validity, accuracy*): es la capacidad que tiene el instrumento de medir lo que se supone que está midiendo.
- Reproducibilidad (en inglés, *reliability, reproducibility*): se refiere a la capacidad del instrumento (o del observador) de producir el mismo resultado cada vez que realiza la medición teniendo los demás factores constantes, es decir, sin que en la realidad haya cambiado el atributo que es medido.

En la **figura 1** se muestra un esquema que explica los conceptos de período de prueba, período de seguimiento, prueba diagnóstica y prueba pronóstica.

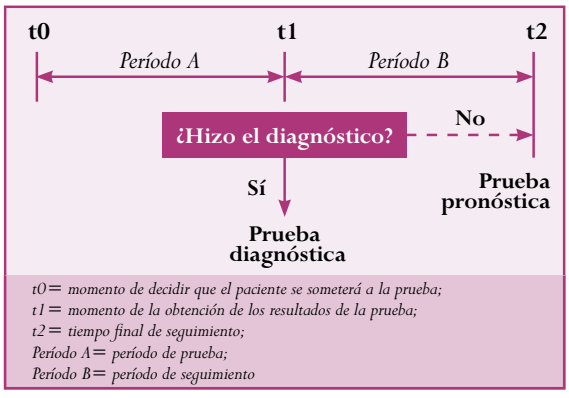

Figura 1. Representación esquemática de los conceptos de prueba diagnóstica y pronóstica

Construcción de la tabla de comparación y cálculo e interpretación de sus celdas internas y externas

La evaluación de tecnologías diagnósticas clásica es la comparación de una prueba que se considera patrón de oro con otra que se está evaluando (prueba en estudio) y de la cual se quiere conocer algunas de sus características para realizar el diagnóstico de la misma enfermedad.

Para los cálculos básicos requeridos en este diseño se crea una tabla de 2 x 2, tal como es mostrada en la **tabla 1**. Esta tabla, también llamada de cuatro celdas, nos permite comparar las dos tecnologías o procedimientos diagnósticos de manera ordenada y con mejor comprensión visual.

Tabla 1. Distribución de los valores en una evaluación de tecnologías diagnósticas

Nueva prueba evaluada	Diagnóstico (patrón)		
	Presente	**Ausente**	**Total**
Positiva	Positivos Verdaderos **(VP)**	Positivos Falsos **(FP)**	Q
Negativa	Negativos Falsos (FN)	Negativos Verdaderos (VN)	1-Q
Total	**P**	**1-P**	**N**

En la parte superior se define el patrón de oro y en la izquierda la prueba que se desea evaluar. Es preferible no memorizar las celdas como a, b, c y d, porque el significado de estas letras cambia si se intercambia la ubicación de la prueba y del patrón de oro (quedando, de esta forma, la prueba por evaluar encabezando las columnas), lo que podría generar confusión. Este problema es similar al que ocurre cuando se analiza un estudio de casos y controles.

Las fórmulas mostradas a continuación indican la forma correcta de calcular los valores de las celdas internas y los subtotales en un estudio de pruebas diagnósticas. Debido a que se calculan proporciones en las fórmulas, los resultados de las expresiones dan un valor entre 0 y 1. Para determinar los porcentajes (%) cada resultado debe multiplicarse por 100.

- Los positivos verdaderos (VP, en inglés *true positive*): son la cantidad de individuos en los que la prueba en estudio es positiva para la enfermedad o desorden y concuerda con el patrón diagnóstico. La expresión matemática de su proporción teniendo en cuenta a todos los individuos del estudio sería la siguiente:

Número de individuos en los que la prueba de estudio es positiva para la enfermedad y concuerda con el patrón de oro

---

Número total de individuos en el estudio

- Los positivos falsos (FP, en inglés *false positive*): son la cantidad de individuos que son positivos para la enfermedad según la prueba de estudio pero en quienes el patrón de oro es negativo (no hay enfermedad). En este caso, la expresión matemática de su proporción sería la siguiente:

Número de individuos en los que la prueba de estudio es positiva para la enfermedad pero el patrón de oro es negativo

---

Número total de individuos en el estudio

- Los negativos falsos (FN, en inglés *false negative*): son la cantidad de individuos que son negativos para la enfermedad según la prueba de estudio, pero que sí la tienen según el patrón de oro. La expresión matemática de su proporción sería la siguiente:

Número de individuos en los que la prueba de estudio es negativa para la enfermedad pero el patrón de oro es positivo

---

Número total de individuos en el estudio

- Los negativos verdaderos (VN, en inglés *true negative*): son la cantidad de individuos en los que la prueba y el patrón de oro son negativos para la enfermedad. Su proporción, expresada como fórmula matemática, sería:

Número de individuos en los que la prueba de estudio es negativa para la enfermedad y concuerda con el patrón de oro

---

Número total de individuos en el estudio

- La prevalencia del diagnóstico (P, en inglés *prevalence of the diagnosis*): es la proporción de individuos que son positivos para el patrón de oro sin importar el resultado de la prueba, es decir, es la suma de las proporciones correspondientes a VP y FN (que equivale al subtotal de la columna izquierda de la **tabla 1**). En términos matemáticos, esta prevalencia es:

Número de individuos en los que el patrón de oro es positivo para la enfermedad

---

Número total de individuos en el estudio

- El complemento de la prevalencia del diagnóstico (1-P): es la proporción de individuos que son negativos para el diagnóstico sin importar el resultado de la prueba, es decir, es la suma de las proporciones que corresponden a FP y VN (en otras palabras, el subtotal de la columna

derecha de la **tabla 1**). Su expresión matemática sería:

Número de individuos en los que el patrón de oro es negativo para la enfermedad

---

Número total de individuos en el estudio

- El nivel de la prueba (Q, en inglés *level of the test*): es la proporción de individuos que son positivos para la prueba sin importar el resultado del diagnóstico, es decir, es la suma de las proporciones correspondientes a VP y FP (en otras palabras, el subtotal de la fila superior de la **tabla 1**). Su representación matemática sería:

Número de individuos con la prueba de estudio positiva para la enfermedad

---

Número total de individuos en el estudio

- El complemento del nivel de la prueba (1-Q): es la proporción de individuos que son negativos para la prueba sin importar el resultado del diagnóstico, es decir, es la suma de las proporciones que corresponden a FN y VN (el subtotal de la fila inferior de la **tabla 1**). Su expresión matemática sería:

Número de individuos en los que la prueba de estudio es negativa para el desorden

---

Número total de individuos en el estudio

- La proporción máxima (uno): se obtiene dividiendo el número total de individuos (N) en sí mismo, equivale al 100% de los individuos del estudio.

Cálculo e interpretación de las medidas utilizadas durante la evaluación de tecnologías diagnósticas

Las medidas de validez que se utilizan con mayor frecuencia son las que se expresan a continuación. Debe aclararse que estas expresiones matemáticas también podrían reemplazarse por el resultado de las ecuaciones anteriores, utilizando las proporciones encontradas y no el número de individuos.

- La sensibilidad *(*S, en inglés *sensitivity*): es la probabilidad de que un individuo tenga una prueba positiva dado que tiene un diagnóstico positivo por el patrón de oro, es decir, es la capacidad de la prueba en evaluación de detectar una enfermedad cuando está presente (observar la columna izquierda de la celda). Su expresión matemática sería:

[Número de individuos VP]

[Número de individuos VP] + [Número de individuos FN]

Y es igual a resolver:

Proporción de positivos verdaderos (VP)

Prevalencia del diagnóstico (P)

- La especificidad *(*E, en inglés *specificity*): es la probabilidad de que un individuo tenga una prueba negativa dado que tiene un diagnóstico negativo por el patrón de oro, es decir, es la capacidad que tiene la prueba de estudio de excluir una enfermedad cuando no existe. En este caso, su expresión matemática sería:

[Número de individuos VN]

[Número de individuos VN] + [Número de individuos FP]

Y es igual a resolver:

Proporción de negativos verdaderos (VN)

Complemento de la prevalencia del diagnóstico (1-P)

- El valor predictivo de una prueba positiva (VPP, en inglés *predictive value of a positive test*): es la probabilidad de que un individuo tenga un diagnóstico positivo por el patrón de oro dado que la prueba fue positiva. Su expresión matemática sería:

[Número de individuos VP]

[Número de individuos VP] + [Número de individuos FP]

Y es igual a resolver:

Proporción de positivos verdaderos (VP)

Nivel de la prueba (Q)

- El valor predictivo de una prueba negativa (VPN, en inglés *predictive value of a negative test*): es la probabilidad de que un individuo tenga un diagnóstico negativo por el patrón de oro dado que la prueba fue negativa a la condición de estudio. Su expresión matemática sería:

[Número de individuos VN]

[Número de individuos VN] + [Número de individuos FN]

Y es igual a escribir:

Proporción de negativos verdaderos (VN)

Complemento del nivel de la prueba (1-Q)

- La eficiencia de la prueba (en inglés *efficiency*): es la probabilidad de que una prueba concuerde con el patrón de oro. Su expresión matemática sería:

[Proporción de VP] + [Proporción de VN]

Ejemplo para la consolidación de los conceptos revisados

Suponga que un investigador quiere determinar la eficacia de un nuevo equipo de ecografía abdominal (prueba de estudio) que se piensa es útil en el diagnóstico de masa renal en un grupo determinado de pacientes menores de cinco años. Esta prueba está siendo comparada con la tomografía axial computada (TAC) que se considera el patrón de oro. Luego de realizar adecuadamente el diseño con muestreo transversal, velar por el cumplimiento de los principios éticos para la investigación en humanos y procesar los datos recolectados, los resultados obtenidos son los siguientes:

- De un total de 1.000 menores estudiados, 600 tienen la TAC (patrón) positiva para masa renal y 500 son negativos al aplicar la ecografía (prueba).
- De los individuos positivos para la ecografía, 400 también eran positivos para la TAC.

Con base en estos escasos datos suministrados y en lo descrito anteriormente deben calcularse todas las expresiones. Para lograr la solución de este ejemplo deben seguirse unos pasos básicos:

1. Dibujar y completar la tabla: el patrón de oro se deja encabezando las columnas. Ver **tabla 2** (los valores dados en el ejemplo se encuentran en negrita, los demás se deducen).
2. Calcular e interpretar cada término:

- VP = 400/1.000 = 0,4*100 = 40%. En el 40% de los menores estudiados tanto la prueba ecográfica como la TAC mostraron masa renal.
- FP = 100/1.000 = 0,1*100 = 10%. En el 10% de los menores la ecografía era positiva pero la TAC era negativa para masa renal.

- FN = 200/1.000 = 0,2*100 = 20%. En el 20% de los menores estudiados la ecografía fue negativa pero la TAC evidenció masa renal.
- VN = 300/1.000 = 0,3*100 = 30%. En el 30% de los menores la ecografía y la TAC fueron negativas para masa renal.
- P = 600/1.000 = 0,6*100 = 60%. El 60% de los menores tuvieron diagnóstico de masa renal por TAC. La prevalencia de masa renal en los menores estudiados fue del 60%.
- 1-P = 400/1.000 = 0,4*100 = 40%. El 40% de los menores no tenían diagnóstico de masa renal.
- Q = 500/1.000 = 0,5*100 = 50%. El 50% de los menores tuvieron evidencia ecográfica de masa renal.
- 1-Q = 500/1.000 = 0,5*100 = 50%. El 50% de los menores no tuvieron evidencia ecográfica de masa renal.
- S = 400/600 = 0,667*100 = 66,7%. La sensibilidad de la nueva ecografía para detectar masas renales en menores de cinco años es del 66,7%. Esto es igual a decir que de 100 menores con diagnóstico de masa renal, es decir, con TAC positivo, 67 fueron también positivos en la ecografía renal.
- E = 300/400 = 0,75*100 = 75%. La especificidad de la nueva ecografía para detectar masas renales en menores de cinco años es del 75%. Esto significa que por cada 100 menores sin diagnóstico de masa renal, es decir, con TAC negativo, 75 fueron correctamente clasificados como negativos en el examen ecográfico.
- VPP= 400/500 = 0,80*100 = 80%. El valor predictivo de la ecografía es del 80% cuando detecta masa renal. Esto es igual a decir que de cada 100 menores con ecografía positiva, en otras palabras, con evidencia ecográfica de masa renal, 80 tenían realmente masa renal (es decir, diagnosticados por TAC).
- VPN= 300/500 = 0,60*100 = 60%. El valor predictivo de la ecografía es del 60% cuando **no** detecta masa renal. Esto es igual a decir que, por cada 100 menores con ecografía negativa, 60 no tenían masa renal.

Tabla 2. Distribución de los menores estudiados según hallazgos ecográficos y tomográficos[a]

Nuevo equipo eco-gráfico	Tomografía renal		
	Masa	**No masa**	**Total**
Positivo	**400**	100	500
	(VP=0,4)	(FP=0,1)	(Q= 0,5)
Negativo	200	300	**500**
	(FN=0,2)	(VN=0,3)	**(1-Q=0,5)**
Total	**600**	400	**1.000**
	(P=0,6)	(1-P=0,4)	**(N= 1,0)**

a. Datos imaginarios. Entre paréntesis se encuentran los valores de las proporciones resultantes al dividir el número de sujetos en cada celda en el número total de sujetos (N = 1.000). Para calcular el porcentaje (%) multiplique estos resultados por 100.

Si el investigador acepta que el nuevo método ecográfico sea distribuido, y teniendo en cuenta una sensibilidad un poco menor al 67%, dejaría escapar a 33 niños y niñas con masa renal, quienes podrían haber sido diagnosticados por TAC. Es necesario tener en cuenta cuál es la finalidad de realizar una prueba diagnóstica a los pacientes: una sensibilidad alta es requerida en pruebas de tamizaje (por ejemplo, la citología del cuello uterino es una prueba rápida y poco costosa en la que se buscan estrategias para tener la menor proporción posible de negativos falsos), mientras que una especificidad alta es requerida para realizar confirmación diagnóstica (por ejemplo, prueba confirmatoria de infección por virus de inmunodeficiencia humana).

En este ejemplo es importante que la masa renal sea observada porque la no detección puede producir un empeoramiento del pronóstico de los menores que tienen una masa secundaria a, por ejemplo, un tumor maligno, y por tanto, este examen debería tener una mayor sensibilidad.

CONCLUSIONES

Una de las decisiones más importantes que deben afrontar los médicos y demás profesionales de la salud durante el ejercicio de la profesión es la de establecer a cuáles herramientas diagnósticas van a someter a sus pacientes ante la sospecha de determinada condición clínica. Esta tarea diaria los obliga a tener claridad en los conceptos y en la forma de interpretar las características que establecen si una prueba es eficaz o no para estudiar la enfermedad que están sospechando.

Como se revisó al inicio del texto, aunque la interpretación puede ser similar, no todos los estudios permiten realizar los cálculos de sensibilidad, especificidad y valores predictivos con las mismas ecuaciones y, por tanto, deben tenerse en cuenta las diferentes formas de realizar el procedimiento de muestreo. En este artículo se evaluaron los términos básicos de un estudio con muestreo transversal y se hizo claridad en su interpretación.

REFERENCIAS

1. Kraemer HC. Evaluating medical tests. Objective and quantitative guidelines. Newbury Park: Sage Publications; 1992.
2. Orozco LC, Camargo DM. Evaluación de tecnologías diagnósticas y tipos de muestreo. Biomédica 1997;17:321-4.
3. Hays RD, Anderson RT, Revicky D. Assessing reliability and validity of measurement in clinical trials. En: Staquet MJ, Hays RD, Fayers PM. Quality of life assessment in clinical trials: methods and practice. Oxford: Oxford University Press; 1998. p. 169-82.
4. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis 1981;34:599-610.
5. Streiner DL, Norman GR. Health measurements scales: a practice guide to their development and use. 2nd ed. Oxford: Oxford Medical Publications; 1995.
6. Singh N, Mishra AK, Shukla MM, Chand SK, Barthi PK. Diagnostic and prognostic utility of an inexpensive rapid on site malaria diagnostic test (ParaHIT f) among ethnic tribal population in areas of high, low and no transmission in central India. BMC Infect Dis 2005;5:50.
7. Tierney MC, Yao C, Kiss A, McDowell I. Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. Neurology 2005;64:1853-9.
8. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13-20.

Conflicto de intereses: no existen compromisos particulares o institucionales por parte de los autores. El ejemplo de pruebas diagnósticas utilizado para consolidar los conceptos explicados en esta publicación es imaginario.