UNIVERSIDAD
NACIONAL
DE COLOMBIA
Revista de la Facultad de Medicina 2001; 49 (3):175-180

Aspectos sobre diseño y tamaño de muestra en estudios de pruebas diagnósticas

Ricardo Sánchez Pedraza, Profesor Asociado, Departamento de Psiquiatría y Centro de Epidemiología Clínica. Jairo Echeverry Raad, Profesor Asociado, Departamento de Pediatría y Centro de Epidemiología Clínica, Facultad de Medicina, Universidad Nacional. Dirección para

SUMMARY

This article deals with general and particular aspects related with the meaning and development of diagnostic tests. Its main subject is about the elements needed for the appropriate sample size in these kind of studies.

RESUMEN

En este artículo se presentan algunas consideraciones generales y particulares con respecto al significado y desarrollo de los estudios sobre pruebas diagnósticas. El centro del mismo se dedica a la formulación de los elementos necesarios que subyacen al cálculo del tamaño de muestra en este tipo particular de diseños.

INTRODUCCIÓN

En Medicina, una Prueba Diagnóstica (Px.Dx.) es cualquier dato que, percibido a través de alguno de nuestros sentidos, puede modificar las probabilidades de un diagnóstico. Dentro de estos términos una Px.Dx. no sólo es aquella que puede ser verificada en un laboratorio, sala de rayos X, o unidad de patología, después de un proceso tecnológico complejo, sino también, mediante lo observado, lo auscultado, lo palpado, lo percutido y demás elementos del ejercicio semiológico cotidiano (1). En este sentido, y como elementos para efectuar un mismo diagnóstico, se pueden tener al alcance, por un lado unas pruebas sofisticadas, costosas e invasivas, y por el otro, sencillas preguntas, la observación o medición del tamaño, color u olor de un individuo o de una parte de su anatomía, la forma de saludar, la descripción de su comportamiento, etc.

En forma habitual las Px.Dx. se han utilizado en el intento de establecer con exactitud y precisión la presencia de alguna enfermedad o condición; sin embargo, también juegan papel importante en el establecimiento de factores de riesgo, estadío o evolución clínica de las enfermedades, respuesta a las terapias e intervenciones, o en la predicción de los desenlaces intermedios o finales (2,3).

Los datos obtenidos para establecer la situación pasada o actual suelen llamarse *pruebas diagnósticas* y los que proyectan sus resultados, con predicción de lo que puede suceder se conocen como *pruebas pronósticas* (4).

El diagnóstico va en un sentido (del clínico hacia su paciente), y la enfermedad en otro (del paciente al médico). Dicho de otra manera, el diagnóstico es lo que uno cree que el paciente tiene y la enfermedad es lo que realmente el individuo tiene. Un diagnóstico es apropiado si ambos sentidos coinciden (4). Dado que esto es un asunto difícil, y a la vez esencia del proceso diagnóstico, el establecimiento de la presencia de una enfermedad se realiza más en términos de probabilidades de su ocurrencia.
El proceso en el que se construye la probabilidad de decirle a alguien que está enfermo o que padece alguna condición, mediante el acopio necesario de Px.Dx. se denomina proceso diagnóstico (1). Diagnosticar significa clasificar a los individuos en alguna condición (5), sea esta una taxonomía nosológica, un síndrome, una enfermedad, el estadío evolutivo, etc.

Cada Px.Dx, sin importar sus características, métodos utilizados, costos o sitio en la historia natural de la enfermedad en la que se ejecute, es susceptible de ser medida en su rendimiento operativo (1).

Características operativas de las pruebas diagnósticas

En el proceso de mejoramiento continuo de calidad, la tecnología de la salud está a la caza de elementos diagnósticos o pronósticos que cada día, con mayor validez y precisión, y menor costo y riesgo, sean más costo-

efectivos en el establecimiento o determinación del estado de las cosas.

Esto se logra mediante el proceso de validación en el que un grupo de individuos se somete tanto a una prueba diagnóstica nueva como ante el estándar de referencia (patrón de oro o "*Gold Standard*"), que no es otra cosa que la mejor prueba diagnóstica disponible para establecer la verdadera condición o situación de un paciente.

Fruto de este proceso de validación se establece para la prueba en cuestión un conjunto de indicadores de exactitud que han sido llamados *Características Operativas de las Pruebas Diagnósticas (COPD).*

Las COPD tradicionales son la *Sensibilidad* (capacidad de la prueba en detectar los enfermos), la *Especificidad* (capacidad de la prueba de detectar a los sanos), el Valor Predictivo Positivo (A la prevalencia del estudio, cuál es la probabilidad de enfermedad si la prueba resultara positiva?), y el *Valor Predictivo Negativo* (A la prevalencia del estudio, cuál es la probabilidad de estar sano si la prueba resultara negativa?).

Últimamente se han utilizado resúmenes de los anteriores indicadores (6,7), como las *razones de verosimilitud* positivas, del inglés *positive likelihood ratios* ó (+) LR, las razones de verosimilitud negativas, del inglés *negative likelihood ratios* ó (-)LR, y más recientemente las razones de ventaja diagnóstica, del inglés *Diagnostic Odds Ratio* (DOR) (2). Estos indicadores se han convertido en medidas de resumen de las COPD, con suma utilidad en el establecimiento más sólido y depurado de las características intrínsecas de ellas, en las selección más apropiada de un punto de corte en la prueba, en la escogencia de la prueba más adecuada entre homólogas o en la determinación de probabilidades pos-prueba (1).

El estudio de una prueba diagnóstica

El diseño general de una prueba diagnóstica es el de un estudio transversal comparativo (8). En general, para su desarrollo se toma un grupo de individuos enfermos y otro de sanos (seleccionados en forma ideal, de un universo, tras un proceso aleatorio), definidos en tales condiciones a la luz de un mismo estándar de referencia. A continuación unos y otros son sometidos, en forma enmascarada e independiente, a la lectura de la prueba bajo investigación (9).

La calidad de un estudio de Px.Dx. se relaciona con aspectos del diseño, con los métodos para reclutar la muestra (en la selección), con la ejecución de las pruebas (en la medición) y con lo completo del reporte del estudio (10,11). Un aspecto trascendente, pero un tanto descuidado a nuestro juicio, como es el número mínimo necesario de pacientes para la obtención de indicadores precisos, será discutido más adelante.

De manera más específica, con respecto a la muestra de pacientes (1,2,4,9,12), esta debe ser seleccionada consecutiva o aleatoriamente, reclutada como una cohorte no clasificada según su estado de enfermedad, en escenarios en donde la prueba bajo estudio pudiera ser realizada en el futuro. El proceso de selección y referencia utilizado y las características demográficas y clínicas de los pacientes deben ser completamente descritas. Es óptimo incorporar en la muestra todo el espectro de enfermedad: Si en la muestra sólo se incluyen individuos severamente enfermos y claramente sanos, la prueba puede no detectar enfermos cuando se aplica en poblaciones en las cuales hay estadíos menos severos de la enfermedad. Esto cobra especial importancia cuando el resultado de la prueba que se estudia se correlaciona con la severidad de la enfermedad; cuando el rendimiento operativo de una prueba se modifica en la medida en que se modifica el espectro de enfermedad, esa prueba revela *Sesgo de Espectro* (13).

En lo referente al diagnóstico por el patrón de referencia (1,2,4,9,14), los métodos y las pruebas deben ser descritos en detalle. El diagnóstico de la condición debe ser el más cercano a la verdad, y el mismo estándar de referencia debe haber sido aplicado a todos los pacientes por igual, desconociendo cualquier dato adicional de ellos, incluso cómo resultaron en la prueba bajo estudio . Es fundamental disponer de un patrón de oro adecuado: Para evaluar una nueva Px.Dx. es necesario que en la muestra se incluyan individuos en los que se haya identificado con buen grado de certeza el desenlace de interés. Si el patrón de oro que me ha definido dicho desenlace es imperfecto se corre el riesgo de calificar como falsos positivos a individuos no detectados por el patrón de oro o como falsos negativos a los erróneamente detectados positivos por el patrón de oro.

En relación con la prueba bajo estudio (1,2,4,9,14), su aplicación debe ser descrita con detalle. A los miembros de la muestra se les deben efectuar las evaluaciones de forma enmascarada, o sea que el investigador que evalúa la prueba debe desconocer el resultado del patrón de oro y el que aplica el patrón de oro debe ignorar el resultado de la prueba que se estudia. Se deben haber reportado todos los resultados incluso aquellos con resultados indeterminados o en "zona gris" y haberse realizado antes de iniciar cualquier tratamiento. Evidencia empírica reciente (16,17), ha revelado que los sesgos más importantes en los estudios de PxDx., radican en el diseño y muy particularmente en los siguientes, en su orden:

i. Reclutar por separado a "enfermos" y a "sanos" (Sesgo de selección).
ii. Dada una información preliminar arrojada por la prueba bajo estudio, ejecutar diferentes patrones de referencia a enfermos y a sanos (Sesgo de medición en la verificación).
iii. Determinar la prueba bajo estudio siendo conocedor (no enmascarado) de la real condición de salud de los individuos, o viceversa, establecer la situación real del individuo mediante

el estándar de referencia conociendo el resultado arrojado por la prueba bajo estudio (Sesgo de medición por un diagnóstico conocido).

Resultados de una prueba diagnóstica

Al culminar el proceso se contará con cuatro posibles pares de resultados inherentes a las pruebas, dos aciertos y dos errores: En los pares de aciertos, la prueba resulta positiva y el individuo se encuentra realmente enfermo (verdadero positivo) o la prueba resulta negativa y el individuo realmente es sano (verdadero negativo). En los errores, la prueba resulta positiva y el individuo se encuentra realmente "sano" o libre de la condición bajo estudio (falso positivo) o la prueba resulta negativa y el individuo realmente está enfermo (falso negativo).

Las observaciones epidemiológicas tienen dos fuentes de error (18-20): El error sistemático y el aleatorio. El error sistemático es la desviación sistemática de la verdad y por ello es llamado también sesgo (21). El sesgo suele introducirse ya sea en la selección de los individuos o en los procesos y elementos para medición de ellos, sean estos los instrumentos de medida o los observadores. Exactitud, ausencia de sesgo o validez, son conceptos similares, aunque no idénticos.

El error aleatorio es aquel que se presenta debido al azar y puede producirse por efectos de la variabilidad biológica del individuo, del observador y de los instrumentos de medida (22). Reproducibilidad, confiabilidad, confianza, consistencia o precisión son palabras que sugieren un escaso error aleatorio.

La ausencia de sesgos no excluye los errores aleatorios y viceversa. Por ejemplo, si un báscula se encuentra descalibrada con un exceso de 200 gramos en cada medida, pero de manera consistente en los mismos sujetos, medida tras medida, los resultados obtenidos muestran poca variabilidad, diríamos que dicha báscula es inexacta pero precisa. El instrumento está sesgado (descalibrado) pero con poco error aleatorio (preciso). El interés epidemiológico es la obtención de indicadores de frecuencia o de efecto que sean ante todo válidos y precisos (23).

Aunque suene reiterativo, si un estudio toma un grupo significativo y suficiente de individuos sanos y enfermos, con un perfil similar al que presenciaríamos en el ejercicio clínico, y a todos los confronta de manera enmascarada e independiente a través tanto de una prueba bajo estudio como del estándar de referencia, diríamos que ese estudio de pruebas diagnósticas es válido (o sea que no se cometieron errores sistemáticos o sesgos en el proceso de selección y medición), y por lo tanto no se desautorizan los resultados obtenidos (22).

Pero lo anterior no anula la posibilidad de que se hayan introducido errores aleatorios en el estudio, y que por esto los resultados puedan haberse presentado sólo por efecto del azar. Este efecto no permite tener la suficiente confianza y precisión en los estimadores del indicador operativo de la prueba diagnóstica en consideración.

No existen pruebas perfectas; cada una posee un grado relativo de aciertos y errores. Una prueba útil será aquella en la que la relación entre los aciertos y sus costos, supere el de los errores y sus costos (6).

Teniendo como referente el diseño general presentado previamente, se pueden plantear diferentes modalidades de estudio de una prueba diagnóstica:

i. Se tiene un patrón de oro robusto y se quieren conocer las características operativas de una nueva prueba, contrastando sus resultados con ese patrón de oro.
ii. Se quieren comparar dos pruebas, una de las cuales tiene características operativas reconocidas.
iii. Se quieren comparar dos pruebas, contrastando su exactitud.
iv. Se desea evaluar la concordancia entre dos pruebas diagnósticas.

El tamaño de la muestra (TM)

Una de las maneras de disminuir la posibilidad de errores aleatorios en un estudio, es calcular el número mínimo necesario de observaciones o individuos para obtener unos resultados precisos y consistentes.

Para resumir, el calcular un tamaño de muestra (TM) permite establecer de manera precisa el verdadero estimador del rendimiento operativo de una prueba diagnóstica o de la diferencia en el funcionamiento de dos pruebas, con un nivel tolerable de error aleatorio. Llama la atención la escasez de literatura alrededor de este tópico, así como la falta de un consenso, evidente en los múltiples enfoques que se plantean en diferentes artículos sobre evaluación de Px.Dx.

Teniendo en cuenta los diversos diseños que pueden plantearse para el estudio de una Px.Dx, no tiene sentido establecer un único método para el cálculo del TM. Por esta razón, ante la pregunta: ¿Qué tamaño de muestra necesito?, la respuesta será: Depende el diseño que tenga el estudio.

Si se tienen en cuenta las características específicas del diseño del estudio, considerando cómo se efectúe el planteamiento de hipótesis, se pueden tener presentes las siguientes opciones:

1. **Se tiene un patrón de oro robusto y se quieren conocer las características operativas de una nueva prueba, contrastando sus resultados con ese patrón de oro:** Teniendo en cuenta que algunos estudios de Px.Dx. con esta estructura pueden semejar un diseño en el cual se analiza una eventual asociación entre una condición (enfermedad o diagnóstico) y un método para

detectarla o medirla (prueba), una aproximación muy general estima que un tamaño debe ser suficiente cuando quiera que hay mínimo 10 individuos en cada casilla marginal de la tabla de 2x2[4]. Estos valores deberán ser ajustados de acuerdo con el número de individuos con y sin la condición que se busca diagnosticar, aspecto que se mostrará posteriormente.

2. Se quiere evaluar una hipótesis sobre características de una nueva prueba: En estos casos lo que se quiere evaluar es si las características de la nueva prueba (sensibilidad, especificidad o valores predictivos) difieren de las de la prueba convencional. Esto se ilustra con el siguiente ejemplo:

Existe una prueba de referencia "A" y después de varios estudios se sabe que su sensibilidad es del 90%. Se desarrolló una nueva prueba "B" cuya sensibilidad se quiere comparar con la de la prueba de referencia. ¿se puede demostrar que es diferente la sensibilidad de la nueva prueba?
De acuerdo con el procedimiento de una prueba de hipótesis, el planteamiento sería el siguiente:
H_0: Sensibilidad de A = Sensibilidad de B
H_a: Sensibilidad de A ≠ Sensibilidad de B

Dichas hipótesis estarán planteadas en términos de proporciones.
Si se considera que los pacientes o las observaciones en las cuales se va a aplicar la nueva prueba son muy similares a los utilizados en otros estudios para evaluar la prueba de referencia, puede no ser necesario volver a aplicar esta última prueba. En tal caso se asume que la sensibilidad de la prueba de referencia "A", expresada como proporción, es conocida y que el objetivo del estudio es estimar la sensibilidad de la nueva prueba "B".

Para el cálculo del tamaño de la muestra se utiliza la siguiente función que, con un nivel de significación a y un poder de 1-ß, permite comparar una proporción conocida π_1 con una proporción π_2 que se va a estimar (24):

$$N = \frac{\{Z_{1-\alpha/2}\sqrt{[\pi_1(1-\pi_1)]} + Z_{1-\beta}\sqrt{[\pi_2(1-\pi_2)]}\}^2}{\delta^2}$$

En esta función π_1 es el valor conocido de la sensibilidad, especificidad o valores predictivos de la prueba patrón, π_2 es el valor que se espera tenga la prueba en dichos parámetros, y $\delta = \pi_2 - \pi_1$.
En el ejemplo que se ha mencionado, si se anticipa que la nueva prueba puede tener una sensibilidad del 95%, los valores dentro de la función para calcular la muestra serían los siguientes:

$\pi_1 = 0.9$
$\pi_2 = 0.95$
$\delta = 0.05$

Los valores de $Z_{1-\alpha/2}$ y $Z_{1-\beta}$ son términos constantes derivados de la distribución normal y se presentan en las tablas 1 y 2. El valor α, también denominado significación o probabilidad de error tipo 1, corresponde a la probabilidad que asignamos al error de rechazar la hipótesis nula cuando esta es cierta. En investigación clínica suele fijarse en el 5% (0.05), aunque, dependiendo del caso, pueden asignarse valores menos estrictos (10% ó 0.1) o más estrictos (1% ó 0.01). El valor ß corresponde a la probabilidad de cometer error tipo 2, es decir de no rechazar la hipótesis nula cuando esta es falsa. Usualmente se establece en un 10 ó 20% (0.1 ó 0.2).

Al reemplazar los valores correspondientes en la función se obtiene un TM de 238 individuos, que es el número de pacientes a los cuales hay que aplicar la nueva prueba para encontrar una diferencia significativa en los valores de sensibilidad.

Tabla 1. *Valores Z para diferentes niveles de alfa, pruebas a dos colas.*

α	α/2	1-α/2	$Z_{1-\alpha/2}$
0.1	0.05	0.95	1.6449
0.05	0.025	0.975	1.96
0.025	0.0125	0.9875	2.2414
0.01	0.005	0.995	2.5758

Tabla 2. *Valores Z para dos niveles de beta.*

ß	1-ß	$Z_{1-\beta}$
0.2	0.8	0.84162
0.1	0.9	1.2816

Dado que en la práctica clínica los valores predictivos dependen fuertemente de la prevalencia de la condición que la prueba mide, se debe hacer la estimación de dichos valores con base en el teorema de Bayes, de la siguiente manera (25):

$$VPP = \frac{Sensibilidad \; x Prevalencia}{\text{Sensibilidad x Prevalencia} + (1\text{-Sensibilidad}) \text{ x } (1\text{-Prevalencia})}$$

$$VPN = \frac{Especificidad \; x \; (1\text{-}Prevalencia)}{\text{Especificidad x } (1\text{-Prevalencia}) + (1\text{-Especificidad}) \text{ x Prevalencia}}$$

El tamaño de la muestra necesario será aquel que resulte mayor luego de calcularlo para los escenarios de sensibilidad, especificidad o valores predictivos, o aquel dependiente de la característica que resulte más importante para la prueba que se está evaluando, por ejemplo la sensibilidad si se trata de una prueba de tamizaje.

3. **Se desea evaluar una hipótesis sobre diferencia de proporciones:** En este caso se aplican simultáneamente dos pruebas diagnósticas a los mismos individuos. Se quiere evaluar la hipótesis nula de no diferencia en las exactitudes que arroja la prueba y el estándar de referencia. Hay que tener en cuenta que en tal situación los datos no son independientes, razón por la cual el cálculo del tamaño de la muestra sobre la premisa de diferencia de proporciones en muestras independientes resulta erróneo.

Cuando se plantean hipótesis sobre diferencia de proporciones en pruebas diagnósticas que se aplican en un mismo individuo, se utiliza la exactitud de la prueba. Recordemos que la *exactitud* es el total de verdaderos positivos y verdaderos negativos que arroja la prueba, y que puede expresarse como una proporción. En tal caso las hipótesis se plantean de la siguiente manera:

H_0: Exactitud de A = Exactitud de B

H_a: Exactitud de A ≠ Exactitud de B

Plantear las hipótesis de esta manera supone que se dispone de un patrón de oro de muy buena calidad, contra el cual se están contrastando las dos pruebas.

La tabla 2x2 correspondiente podría tener esta estructura:

Resultados exactos*

	Si	No
Prueba A		
Prueba B		

*Resultados exactos: Corresponden a la suma de verdaderos positivos y negativos. En cada celda de la columna correspondiente a "Si", se incluye el número de casos verdaderos positivos y verdaderos negativos para cada una de las pruebas.

Teniendo en cuenta la situación de no independencia, la anterior tabla 2x2 puede construirse de tal modo que las unidades que queden en cada celda sean los pares de observaciones efectuados sobre un mismo individuo. El diseño de la tabla es ahora el siguiente:

Prueba A
Resultados exactos

Prueba B Resultados exactos	Si	No	Total
Si	r	s	r + s
No	t	u	t + u
Total	r + t	s + u	N pares

Esta tabla tiene la estructura utilizada en la prueba de McNemar (26). En este tipo de estadístico se trabaja con los denominados pares discordantes, que son los correspondientes a las celdas s y t, ya que los pares concordantes (celdas r y u) no dan información sobre las diferencias que se están evaluando (27). La proporción de discordantes $(\pi Discordantes)$ puede calcularse así:

$$\pi \text{ Discordante} = \frac{(s+t)}{N \text{ pares}}$$

El cociente ψ dice qué tanto mayor es la discordancia de la prueba B en relación con la discordancia de la prueba A. En la tabla de pares equivale a dividir s entre t:

$$\psi = s/t$$

Para calcular el tamaño de la muestra es necesario anticipar la proporción de discordancia entre las dos pruebas y el cociente ψ. La función que se aplica es la siguiente:

$$N=\frac{\{Z_{1-\alpha/2}(\psi+1)+Z_{1-\beta}\sqrt{[(\psi+1)^2-(\psi-1)^2\pi_{Discordancia}]}\}^2}{(\psi-1)^2\pi^2_{Discordancia}}$$

Si se quieren comparar dos pruebas con una probabilidad de discordancia de 0.15, teniendo una discordancia tres veces mayor en una prueba que en otra, se tienen los siguientes valores para reemplazar en la función:

$$\pi_{Discordancia} = 0.15$$
$$\psi = 3$$
$$Z_{1-\alpha/2} = 1.96$$
$$Z_{1-\beta} = 0.841$$

Con los anteriores valores se requiere una muestra de 207 individuos a los cuales aplicar las dos pruebas.

4 Se quiere evaluar una hipótesis de concordancia: En tal caso se asume que hay N sujetos evaluados por dos pruebas diferentes. Así, la probabilidad estimada de desacuerdo entre las dos pruebas es:

$$\pi_{Dis} = d \div N$$

Donde d es el número de desacuerdos entre las dos pruebas. Como puede verse, aquí se aplica un criterio similar al del caso en el que se plantea la evaluación de diferencia de proporciones.

Para una amplitud determinada $w\pi$ de un intervalo de confianza al $100(1\alpha)\%$, el tamaño de la muestra será:

$$N=\frac{4\,\pi\,Dis\,(1-\pi\,Dis)\,Z^2_{1-a/2}}{W^2\pi}$$

¿Cuántos enfermos y cuántos sanos?

Para definir el número de pacientes en la muestra con y sin la condición que se está midiendo, el criterio es incluir un número similar de pacientes dentro de cada uno de los posibles espectros de la enfermedad, esto es, intentar en lo posible reproducir la prevalencia o todo el espectro de la enfermedad, similar al que se verificaría en la práctica clínica. Por ejemplo, si la enfermedad tiene tres niveles de severidad (leve, moderado y grave), y el tamaño calculado de muestra son 400 pacientes, se deberían incluir 100 pacientes sin la enfermedad, 100 leves, 100 moderados y 100 severos.

Otro criterio, fundamental cuando se evalúan características operativas, es que la proporción de enfermos y sanos en muestra refleje la prevalencia de la condición que se busca diagnosticar, en la población en la cual se aplicará la prueba. En general, es recomendable incluir más individuos enfermos que sanos, aunque se estén creando artificialmente prevalencias superiores al 50%, ya que de esta manera se tendría más oportunidad de incluir pacientes con espectros de enfermedad desconocidos o posiblemente no detectables.

REFERENCIAS BIBLIOGRÁFICAS

1. **Echeverry J, Ardila E.** Pruebas diagnósticas y proceso diagnóstico. En: Ardila E, Sánchez R, Echeverry J. Estrategias de Investigación en Medicina Clínica. Bogotá: Editorial Manual Moderno: 2001.
2. **Deeks JJ.** Systematic reviews on evaluations of diagnostic and screening test. BMJ 2001;323:157-162.
3. **Ruiz A, Ruiz JG.** Fundamentos de Investigación Clínica. Exámenes diagnósticos: Aproximación a su uso racional. Pediatría 1993;28(2):111-119.
4. **Kramer HC.** Medical Test: Objective and Quantitative Guidelines. SAGE Publications, Inc. Newburry Park, Cal, USA. 1992. Chapter: 14.
5. **De Almeida N.** Epidemiología sin Números: Una introducción crítica a la ciencia epidemiológica. Serie Paltex N.28. OMS-OMS. 1992.
6. **Suchman AL, Dolan JG.** Odds and Likelihood Ratio. In: Griner, Panzer RJ, Greenland P, eds. Clinical diagnosis and the laboratory. Logical strategies for common medical problems. Chicago: Year Book Medical Publisher., 1986:36-43.
7. **Zweig MH, Campbell G.** Receiver - Operating Characteristic Curve (ROC) Plots: A fundamental Evaluation Tool in Clinical Medicine. Clin Chem 1993;39(4):561-77.
8. **López F.** Interpretación de Pruebas Diagnósticas. En: López F. Manual de Medicina Basada en la Evidencia. Editorial Manual Moderno & JGH Editores, 2000.
9. **Dawson-Sanders B, Trapp RG.** Procedimientos para valorar diagnósticos. En: Dawson-Sanders B, Trapp RG. Bioestadística Médica. Manual Moderno, México 1990:265 -82.
10. Cochrane Methods Group on Systematic Review of Screening and Diagnostic Test. Recommended methods [update 6 jun 1986]. www.cochrane.org/cochrane/sadtdoc1.htm
11. **Reid MC, Lachs MS, Feinstein AR.** Use of methodological standards in diagnostic test research. Getting better but still no good. JAMA 1995;274:645 - 651.
12. **Knapp RG, Miller III MC.** Clinical Epidemiology and Biostatistics. Baltimore:Williams & Wilkins: 1992. p 42-43.
13. **Lachs MS, Nachamkin I, Edelstein PH, et al.** Spectrum Bias in the Evaluation of a Diagnostic Tests: Lessons from the rapid Dipstick Test for Urinary Tract infection. Ann Int Med 1992;117:135 - 140.
14. **Knapp RG, Miller MC.** Describing the Performance of a Diagnostic Test In:. Knapp RG, Miller MC. Clinical epidemiology and biostatistics. Malvern (Pennsylvania): Harwal Publishing Company, Baltimore,Williams & Wilkins: 1992: 31-51.
15. **Bates AS, Margolis PA, Evans A.** Verification bias in paediatric studies evaluating diagnostic tests J Pediatr 1993;122: 585 -90.
16. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Maeulen JHP et al.** Empirical evidence of design-related bias in studies of diagnostic test. JAMA 1999;282:1061-66.
17. **Begg CB.** Biases in the assessment of the diagnostic tests. Stat in Med 1987;6:411 -23.
18. **Knapp RG, Miller MC.** Describing the Performance of a Diagnostic Test In:. Knapp RG, Miller MC. Clinical epidemiology and biostatistics. Malvern (Pennsylvania): Harwal Publishing Company, 1992: 31-51.
19. **Sox HC, Blatt MA, Higgins MC, Marton KI.** Probability: Quantifying Uncertainty. In: Sox HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Boston: Butterworth Publishers, 1988:103 -46.
20. **Riegelman RK. Hirsch RP.** Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura biomédica. 2ª. Ed. Washington, DC. : OPS, 1992. Publicación científica: 531.
21. **Ransohoff DF, Feistein AR.** Problems of spectrum and bias in evaluating the efficacy of diagnostic test. N Eng J Med1978;299:926 -30.
22. **Rothman KJ, Greenland S.** Precision and validity in a Epidemiologic Studies. In: Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Lippincott - Raven publishers, Philadelphia, 1998:115- 134.
23. **Rotmnan KJ, Greenland S.** Precision and validity in Epidemiologic Studies. In: Modern Epidemiology Rotmnan KJ, Greenland S. 2ª Edition. Lippincott-Raven publishers. Philadelphia. 1998:115-34.
24. **Machin D, Campbell M, Fayers P, Pinol A.** Sample size tables for clinical studies, 2nd ed.Oxford: Blackwell Science: 1997. p. 21.
25. **Tobias A.** Summary statistics report for diagnostic tests. Stata Technical Bulletin 2000;56: 16-18.
26. **Norman GR, Streiner DL.** Bioestadística. Madrid: Mosby/Doyma Libros: 1996. pp. 155.
27. **Rosner B.** Fundamentals of Biostatistics. 4th ed. Belmont: Duxbury Press: 1995. pp. 378-379.
28. **Machin D, Campbell M, Fayers P, Pinol A.** Sample size tables for clinical studies, 2nd ed.Oxford: Blackwell Science: 1997. p. 71.

Correspondencia: Ricardo Sánchez Pedraza (risanche@bacata.usc.unal.edu.co), Jairo Echeverry Raad (jechever@bacata.usa.unal.edu.co); Centro de Epidemiología Clínica, Instituto Materno Infantil