Colegio Mexicano de Inmunología Clínica y Alergia, A. C. CMICA
Revista Alergia México raM



# Biases in diagnostic test studies: impact on estimating sensitivity and specificity

# Sesgos en los estudios de pruebas de diagnóstico: implicación en la estimación de la sensibilidad y especificidad

Mario Enrique Rendón-Macías,[1] María Valenzuela,[1] Miguel Ángel Villasís-Keever[2]

## Abstract

Diagnostic tests make it possible to determine whether a person has a disease or not. Before incorporating a new diagnostic test in the clinical setting, it is necessary to define its validity through its indicators of performance, sensitivity, and specificity. In these studies, like in any research, the results might not be reliable when there are biases during their execution. This article entails the discussion about the biases in diagnostic test studies that may cause inaccuracy in sensitivity and specificity. The main biases that affect the validity of these studies are: incorporation bias, partial and/or differential verification bias, an imperfect reference standard, a limited spectrum of the disease, and the ambiguous results of the test to be validated. In addition, examples of how these biases impact on the results of sensitivity and specificity are given in this paper.

Key words: Sensitivity; Specificity, Bias; Diagnostic tests

Este artículo debe citarse como: Rendón-Macías ME, Valenzuela M, Villasís-Keever MA. Sesgos en los estudios de pruebas de diagnóstico: implicación en la estimación de sensibilidad y especificidad. Rev Alerg Mex. 2020;67(2):165-173

ORCID
Mario Enrique Rendón-Macías, 0000-0001-7310-6656; María Valenzuela, 0000-0001-6225-6835;
Miguel Ángel Villasís-Keever, 0000-0002-8566-0811

[1]Universidad Panamericana, Escuela de Medicina, Ciudad de México, México
[2]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Unidad de Análisis y Síntesis de la Evidencia, Ciudad de México, México

Correspondencia. Mario Enrique Rendón-Macías.
mrendon@up.edu.mx

Recibido: 2020-06-15
Aceptado: 2020-06-29
DOI: 10.29262/ram.v67i2.771

CC BY NC

## Resumen

Las pruebas de diagnóstico permiten determinar si una persona tiene o no una enfermedad. Para incorporar una nueva prueba de diagnóstico en el ámbito clínico primero es necesario definir su validez a través de sus indicadores de desempeño, sensibilidad y especificidad. Como en cualquier investigación, en este tipo de estudios es posible que los resultados no sean confiables cuando hay sesgos durante su ejecución. En este artículo se discuten los sesgos en estudios de prueba diagnóstica que pueden ocasionar que la sensibilidad y especificidad no sean correctas. Los principales sesgos que afectan la validez en estos estudios son el sesgo de incorporación, la verificación parcial o diferencial, un estándar de oro imperfecto, un espectro limitado de la enfermedad y los resultados ambiguos de la prueba por validar. Además, en este artículo se dan ejemplos de cómo impactan estos sesgos en la sensibilidad y especificidad.


Palabras clave: Sensibilidad; Especificidad; Sesgos; Pruebas diagnósticas

## Abreviaturas y siglas

EO, estándar de oro
NPD, nueva prueba de diagnóstico

## Antecedentes

El avance de los conocimientos e incorporación de nuevas tecnologías en salud se traduce, en parte, en la incorporación de nuevas pruebas diagnósticas, con lo cual se pretende mejorar la calidad de la atención a la salud, o bien, hacer más eficientes los costos.[1,2] Cada año surge un gran número de estudios que se realizan tanto para validar una nueva prueba de diagnóstico (NPD), como para determinar si es más eficiente que las pruebas utilizadas cotidianamente, en términos de velocidad para su ejecución o para causar menor molestias o daños secundarios.[3]

En general, una prueba diagnóstica sirve para identificar si un paciente o un grupo de pacientes tiene una enfermedad determinada o, por el contrario, se descarta dicha enfermedad. Pero, además, las pruebas diagnósticas coadyuvan a definir las causas o mecanismos implicados en la enfermedad de un paciente, así como para definir si un paciente ha desarrollado una o más complicaciones en una enfermedad, decidir posibles acciones terapéuticas (con base en los cambios de un marcador clínico, bioquímico o de imagen), o bien, para estimar el pronóstico.[1,2,4]

Para aceptar que una NPD puede ser útil, primero se debe demostrar su validez. El principal criterio de validez de una NPD es determinar su desempeño, es decir, su capacidad para clasificar correctamente si un paciente tiene (sensibilidad) o no (especificidad) la enfermedad en estudio.[5,6] Posterior a conocer el desempeñó de un NPD, se podrán hacer evaluaciones para conocer la seguridad de la prueba, su facilidad de aplicación o elaboración, rapidez de realización y los costos de su ejecución.

## Diseño de un estudio de prueba diagnóstica

Dada la importancia de determinar la validez de una NPD (también denominada *prueba índice*), cuando se planea la realización de un estudio para su evaluación, los investigadores deberán llevar a cabo una metodología estricta. Así, un buen estudio de prueba diagnóstica debe considerar que la comparación de los resultados (positivos o negativos a la enfermedad) es con un *estándar de oro* (EO), el cual corresponde a la prueba que, en el momento del estudio, se conoce como la mejor para confirmar o descartar un diagnóstico determinado.[5,7]

Además, el estudio debe ser *cegado*, es decir, los investigadores que interpretan la NPD deben desconocer el resultado obtenido del EO y viceversa. Otro aspecto importante en la planeación de un estudio de prueba diagnóstica es el grupo de participantes; al menos, se deberá incluir un grupo de pacientes que tengan la condición en estudio y otro sin dicha con-

dición. Sin embargo, es mejor un estudio con todo el *espectro de la enfermedad*, lo que significa incluir pacientes con diferentes grados de la enfermedad, desde aquellos con síntomas leves, hasta los que obviamente tienen la enfermedad en estudio e, incluso, pacientes graves.[5,6]

Por último, es necesario que el EO y la NPD se realicen al mismo tiempo o en un tiempo corto entre la realización de una y otra prueba, a fin de evitar modificación de las condiciones clínicas de los participantes, dado que la enfermedad puede progresar. De esta forma, es posible que la prueba índice no clasifique como enfermo a un participante en un momento temprano de la enfermedad y, tiempo después, al realizar el EO se confirme el diagnóstico cuando la enfermedad ya está avanzada. Lo anterior se denomina *falso-negativo*.[5,6]

## Sesgos en los estudios de prueba diagnóstica

Los errores que ocurren sistemáticamente en la ejecución en un estudio de investigación se denominan sesgos, los cuales causarán que los resultados de dicha investigación no sean válidos.[8] Para los propósitos de este artículo nos enfocaremos a los problemas metodológicos y las consecuencias en los resultados obtenidos de un estudio en el cual se desea determinar la valides de una NPD. En general, cuando hay sesgos en los estudios de evaluación de pruebas diagnóstica, los resultados obtenidos afectan directamente al desempeño de la NPD en cuanto a sensibilidad y especificidad, haciendo que uno o los dos estén sobreestimados o subestimados.[1]

Si bien, el resultado de una NPD puede ser en un rango de valores, en este artículo solamente consideraremos un resultado dicotómico: afirmativo (con la enfermedad) o negativo (sin la enfermedad). La mejor NPD es la que clasifica apropiadamente a todos o la mayoría de los participantes, en concordancia a lo que indica el EO, ya sea al confirmar o descartar la condición. El porcentaje de acierto (o capacidad de la NPD para detectar a los positivos con enfermedad) se le conoce como sensibilidad, la cual va de 0 % (nunca acierta) a 100 % (nunca se equivoca). A su vez, la prueba también deberá clasificar como negativos a todos los sujetos sin la enfermedad de interés. El porcentaje de acuerdo con el EO al descartar la enfermedad se conoce como especificidad, que también va de 0 a 100 %.[9]

Cuando un estudio de prueba diagnóstica se ejecuta con la suficiente calidad metodológica o sin sesgos, entonces los resultados de sensibilidad y especificidad de una NPD serán considerados como verdaderos. Si bien, se han descrito más de 14 sesgos en los estudios de pruebas diagnósticas;[5,6,7] los que principalmente impactan en la evaluación del desempeño de la NPD se señalan a continuación:[10,11,12]

- *Sesgo de incorporación*: parte de los componentes de la NPD está incluida en los criterios del EO para definir la condición del paciente.
- *Sesgo de verificación parcial:* la realización del EO no se lleva a cabo al total de los pacientes a quienes se realizó la NPD; por ejemplo, el EO se realiza únicamente a quienes tuvieron un resultado positivo en la NPD.
- *Sesgo de verificación diferencial:* sucede cuando no existe o no se puede disponer del EO aceptado en el mundo, por tanto, se recurre a otras pruebas que se aproximan a los resultados del EO. De esta forma, es alta la posibilidad de clasificar erróneamente si un paciente tiene o no la condición en estudio.
- *Sesgo del EO imperfecto:* surge cuando el EO aceptado en el mundo puede clasificar mal a los pacientes. Esto quiere decir que, la prueba considerada como EO puede definir a un paciente como enfermo cuando no lo es y viceversa.
- *Sesgo del espectro de la enfermedad:* ocurre cuando no se incluyen pacientes en todas las condiciones de gravedad de la enfermedad en cuestión.
- *Sesgo de resultados ambiguos o no concluyentes:* aparece cuando los resultados de la NPD no determinan certeramente si un paciente es positivo o negativo, por lo que los investigadores los eliminan del análisis.

## Impacto de los sesgos en los resultados de un estudio de prueba diagnóstica

Para mostrar cómo los sesgos de pueden modificar la estimación de la sensibilidad y especificidad, utilizaremos los resultados de un estudio hipotético.

En primer lugar presentamos los "resultados verdaderos", es decir, sin sesgos. La figura 1 muestra los resultados del desempeño de una NPD, la cual se contrastó con un EO altamente confiable.

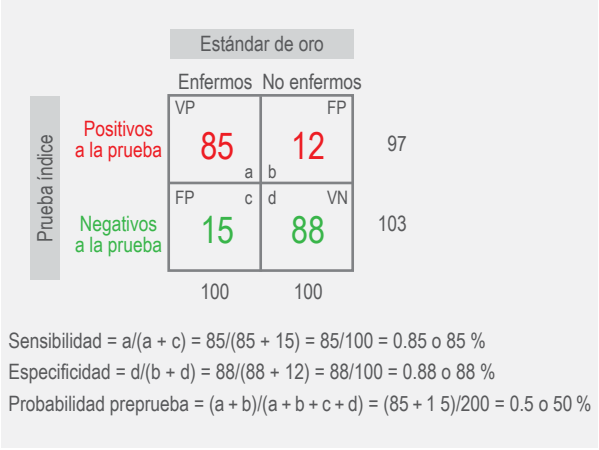


Figura 1. Estudios hipotéticos con "resultados verdaderos". VP = verdaderos positivos, FP = falsos positivos, VN = verdaderos negativos, FN = falsos negativos.

En el estudio se incluyen 200 individuos, 100 con la enfermedad y 100 sin la enfermedad, estos últimos se pueden considerar como grupo control. Como se observa, la NPD tiene una sensibilidad de 85 % y una especificidad de 88 %.

El posible efecto del *sesgo de incorporación* se observa en la figura 2, en la cual se incrementó (sobreestimación) la sensibilidad y la especificidad. En este tipo de sesgos, los autores al conocer previamente la condición clínica de los pacientes pudieran ser más benévolos en ofrecer una prueba positiva en los enfermos y ser más estricto para el grupo de los controles. Con lo anterior incrementan los verdaderos positivos (casilla a) y, por tanto, la sensibilidad; así mismo incorporan a más individuos falsos positivos (casilla b) en el grupo de los verdaderos negativos (casilla d) y, por tanto, también aumentan la especificidad. El resultado sería una prueba reportada falsamente como de alta validez, ya que su desempeño es muy parecido al del EO.

En la figura 3 se muestra cómo se podrían modificar los resultados ante un *sesgo de verificación parcial*. En este caso, los 97 pacientes que salen positivos a la NPD son los únicos sometidos al EO. Por consiguiente, no hay individuos en la casilla "c". Si se consideran a los "ausentes" con posibles resultados negativos, se estimaría una sensibilidad de 100 %. Por otro lado, la especificidad también se incrementa (aunque sea poco), en razón directa con el número de individuos negativos a la NPD no evaluados.

Para el *sesgo por un EO imperfecto* (figura 4), existen tres posibilidades: la primera es una mala clasificación tanto para los enfermos como los sa-

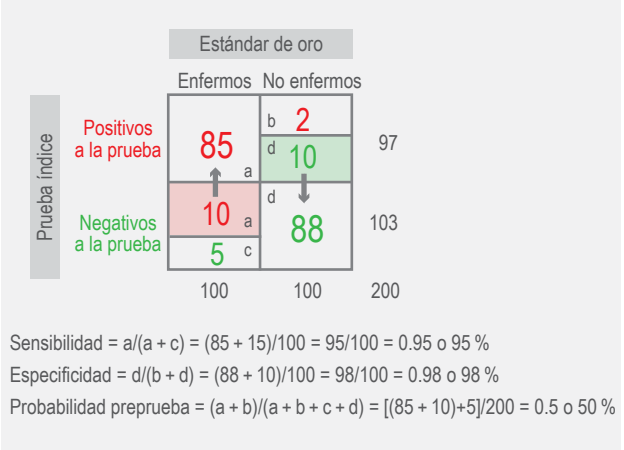


Figura 2. Resultado posible en un sesgo de incorporación.

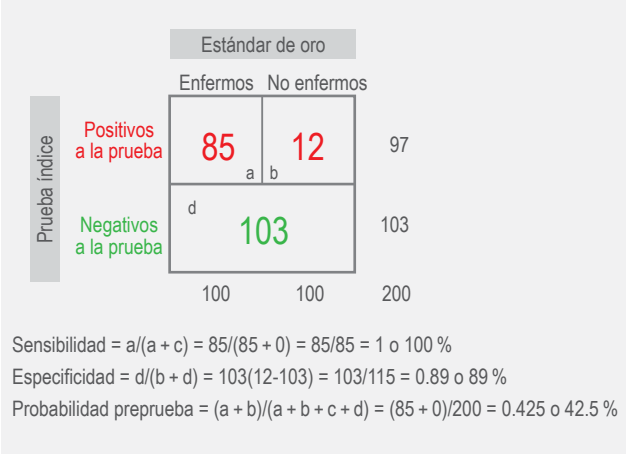


Figura 3. Resultado posible con un sesgo de verificación parcial.

nos. Bajo esta premisa, y dependiendo del número de individuos mal clasificados, la sensibilidad y la especificidad se modificarán. En el ejemplo, ocho individuos enfermos son clasificados como no enfermos, por lo que al tener una NPD positiva pasarán al grupo de falsos positivos (casilla b). De la misma forma, un paciente no enfermo con NPD positiva pasará a al grupo de verdaderos positivos. Por otro lado, dos pacientes falsos negativos (casilla c) pasan al grupo de verdaderos negativos (casilla d) y nueve verdaderos negativos (casilla d) serán incorporados al grupo de falsos negativos (casilla c). Con estos cambios, la sensibilidad y la especificidad se reducen (subestimación). En este tipo de sesgo, los cambios de sensibilidad y especificidad son impredecibles, ya que depende del número de individuos mal clasificados.

La segunda posibilidad del *sesgo por un EO imperfecto* es que la mala clasificación sea unidireccional; el error es más predecible con un EO sesgado solo a los enfermos, es decir, pudiera no detectar a todos, pero siempre clasifica correctamente a quienes no tienen la enfermedad (por ejemplo, resultados de estudios histopatológicos con biopsias mal tomadas). Como se muestra en la figura 5, la sensibilidad no se modifica, pero la especificidad disminuye al existir mayor posibilidad de falsos positivos. La tercera posibilidad, por *sesgo por un EO imperfecto* es contraria al ejemplo previo. El EO está sesgado para la clasificación solo de los no enfermos o grupo control, es decir, identificarlos como enfermos, sin embargo, ningún enfermo se cataloga como no enfermo. La especificidad se mantiene sin cambios, pero la sensibilidad se reduce porque aumentan los falsos negativos (figura 6).

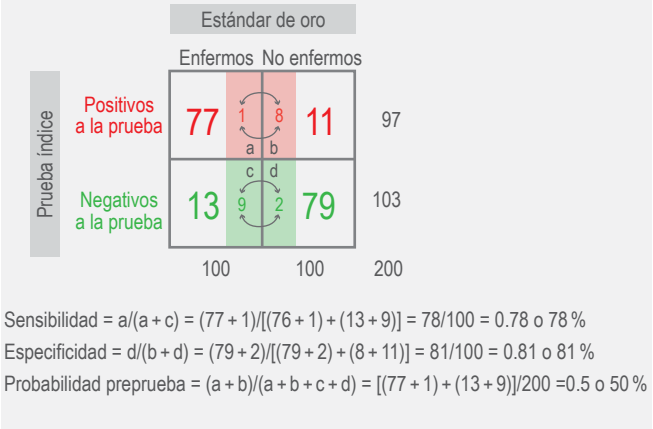


Figura 4. Resultado posible con un sesgo de estándar de oro impreciso (bidireccional).

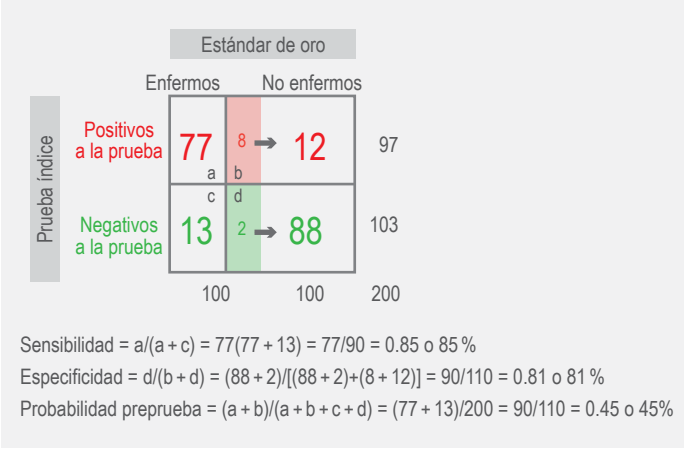


Figura 5. Resultado posible con un sesgo de estándar de oro impreciso, con error en la clasificación de los enfermos (unidireccional).

Por otro lado, cuando se presenta *sesgo de espectro reducido o de extremos* de una enfermedad, los autores solo incluyen en el estudio a pacientes que claramente tienen la enfermedad y los comparan con un grupo de sujetos sanos. En la figura 7 se ejemplifica esta situación, que da lugar a una sobreestimación de la sensibilidad y especificidad. Por el contrario, si se hubieran incluido solo los casos leves y los controles con enfermedades similares, la sensibilidad y la especificidad se reducirían (ver área gris, en la que la sensibilidad es de 84 % y la especificidad de 68 %).

El último ejemplo es el *sesgo de resultados ambiguos* (figura 8). El resultado está sesgado porque al incluir únicamente a participantes con datos certeros de la NPD se modifican los resultados de la sensibilidad y especificidad.

## Análisis de sesgos en estudios de prueba diagnóstica publicados

Finalmente, para demostrar que en estudios publicados los sesgos pueden modificar la interpretación de resultados agregamos dos ejemplos.

### Estudio 1

En el trabajo de Ramírez Enríquez *et al*.[13] el objetivo fue evaluar la validez de la IgE sérica total para el diagnóstico de alergia en una población pediátrica. Los autores revisaron 248 expedientes de pacientes de la consulta externa cuyo número de registro fuera non y contarán con una medición de IgE sérica total y pruebas cutáneas (EO). El estudio puede tener un sesgo de EO imperfecto unidireccional. Las pruebas cutáneas pueden clasificar con error a los enfermos; un resultado positivo asegura clasificar como tal a un

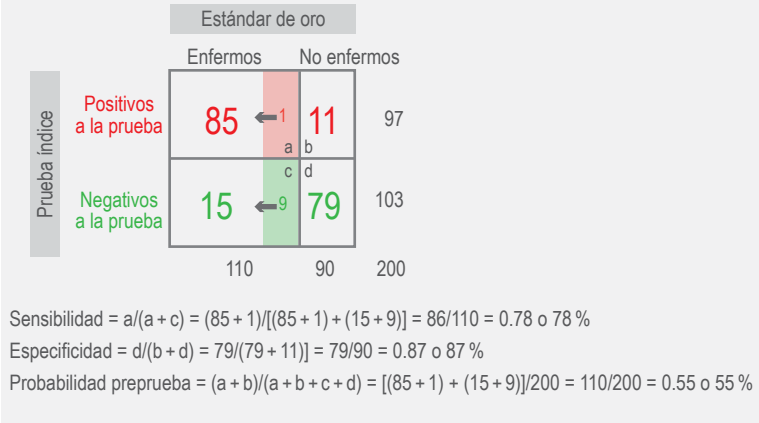


Figura 6. Resultado posible con un sesgo de estándar de oro impreciso, con error en la clasificación de los no enfermos (unidireccional).

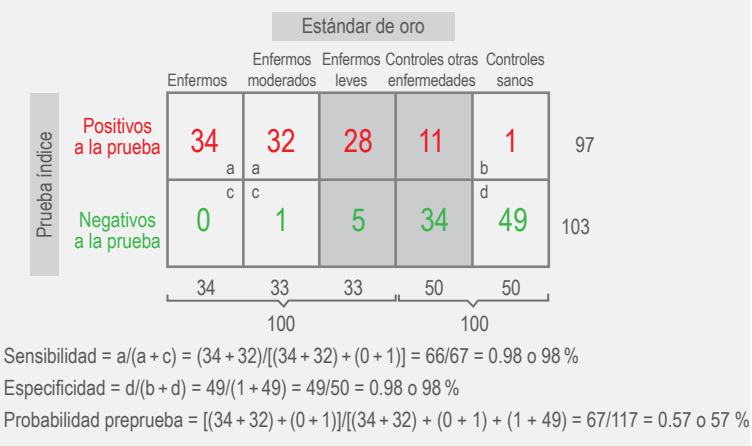


Figura 7. Resultado posible con un sesgo de "espectro reducido de la enfermedad".

enfermo, pero una negativa no lo descarta. La prueba cutánea depende del alérgeno usado y del tipo de síntomas del paciente. Para los no enfermos, la prueba siempre es negativa. Por lo anterior, dado este sesgo, es posible estimar una sensibilidad adecuada, pero existe el riesgo de que la especificidad se subestime. Otro posible sesgo es el de espectro reducido de la enfermedad. Los pacientes seleccionados presentaban diferentes tipos de alergias (respiratoria, gastrointestinal y cutánea) y no se informa la severidad de los síntomas. Aunque el muestreo fue pseudoaleatorio, es probable que solo se incluyeran pacientes con síntomas importantes para ser revisados en una clínica, por lo que los pacientes con síntomas leves pudieron estar subrepresentados. Así, la estimación de la sensibilidad y especificidad de la IgE pudiera ser sesgadamente más alta. Finalmente, los autores no especifican claramente cómo consideraron a los pacientes con valores de IgE justo en los puntos de corte dado: indican como positivos a quienes tuvieron niveles > 15 UI en menores de un año, > 60 UI en los de uno a cinco años, > 90 UI en los de seis a nueve años y > 200 UI en los de 10 a 15 años; para los negativos se consideró < 15, < 60, < 90 y < 200, respectivamente. Si en los estudios no hubo pacientes con alguno de estos valores, entonces potencialmente existe sesgo de resultados ambiguos. Finalmente, debemos señalar que no se detectó sesgo de incorporación, ya que la NPD no está incluida en el EO.

## Estudio 2

En el estudio de Marraccinia *et al*.,[14] en el que se evaluó la validez de la prueba de "activación de basófilos séricos" (NPD) para confirmar la hipersensibilidad a

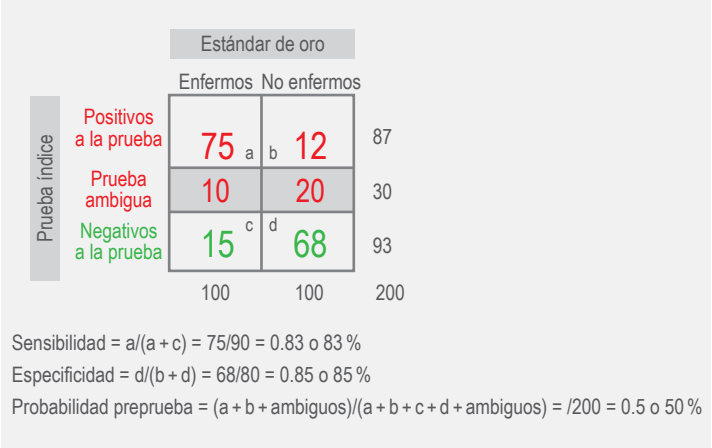


Figura 8. Resultado posible con un sesgo de prueba diagnóstica con resultados ambiguos.

medicamentos en 204 pacientes con antecedentes de reacción adversa. Los autores recurrieron a diferentes pruebas cutáneas, intradérmicas y ensayos de IgE específicas para determinar hipersensibilidad, debido a que la prueba de reto no se pudo realizar en todos los pacientes (en algunos por riesgo de anafilaxia), que es el EO para una respuesta de hipersensibilidad a un medicamento.

Con lo anterior, el sesgo de verificación parcial llevó a no poder lograr un buen estimado de la validez de la prueba de activación de basófilos. Al analizar solo a los pacientes con prueba de reto, su muestra se redujo significativamente, por lo que la estimación de sensibilidad y especificidad realmente no parece ser la reportada. Si analizan los resultados mediante un constructo con todas las pruebas, la posibilidad de un sesgo de EO imperfecto era altamente posible, al no clasificarse adecuadamente a los pacientes con y sin hipersensibilidad al medicamento. En este artículo no encontramos la presencia de un sesgo de incorporación, ya que la NPD no estaba incluida en las consideradas para la clasificación de los pacientes y, al parecer, fue realizada antes del establecimiento del diagnóstico.

Por otro lado, Marraccinia *et al*. incluyeron pacientes con todo el espectro de la enfermedad, desde pacientes con anafilaxia hasta pacientes con cuadros más leves caracterizados por urticaria o síntomas gastrointestinales leves, ante lo cual es poco probable un sesgo de espectro de la enfermedad. Finalmente, los autores no mencionan si existieron resultados ambiguos de la NPD, a fin de descartar un sesgo de espectro de la prueba.

Como se puede apreciar en los ejemplos, los sesgos incrementan que los valores estimados de sensibilidad y especificidad no sean los reales. Por lo anterior, consideramos que es importante para un lector crítico conocer de su existencia. Si bien, es posible que los sesgos sean cometidos involuntariamente, por desconocimiento o falta de cuidado en la realización de un estudio, es necesario recordar que los sesgos metodológicos no pueden ser corregidos con procedimientos estadísticos, dado que estos últimos solo permiten estimar el grado de variación aleatoria causado por el tamaño de la muestra utilizado.[2,3,5]

Finalmente, para los estudios de validación de NPD con resultados no dicotómicos es necesario también informar acerca de los criterios para establecer la existencia de uno o más puntos de corte para la toma de decisiones.

En conclusión, la validez de un estudio de prueba diagnóstica básicamente está determinada por que no se cometan sesgos durante su ejecución. Cuando ocurren sesgos, entonces los resultados obtenidos de una NPD, en términos de sensibilidad y especificidad, pueden estar sobre o subestimados. Los principales sesgos que impactan en los resultados de una prueba diagnóstica son de incorporación, verificación parcial o diferencial, EO imperfecto, espectro inadecuado y cuando los resultados de la NPD son ambiguos.

## Referencias


1. Chassé M, Fergusson DA. Diagnostic accuracy studies. Semin Nucl Med. 2019;49(2):87-93. DOI: 10.1053/j.semnuclmed.2018.11.005
2. Escrig-Sos J, Martínez-Ramos D, Manuel Miralles-Tena J. Pruebas diagnósticas: nociones básicas para su correcta interpretación y uso. Cir Esp. 2006;79(5):267-273. DOI: 10.1016/S0009-739X(06)70871-5
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BJM. 2003;326(7379):41-44. DOI: 10.1136/bmj.326.7379.41
4. Villasís-Keever MÁ, Miranda-Novales MG. El protocolo de investigación II: los diseños de estudio para investigación clínica. Rev Alerg Mex. 2016;63(1):80-90. DOI: 10.29262/ram.v63i1.163
5. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189-202. DOI: 10.7326/0003-4819-140-3-200402030-00010
6. Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013;66(10):1093-1104. DOI: 10.1016/j.jclinepi.2013.05.014

7. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard-an update. PLoS One. 2019;14(10):e0223832. DOI: 10.1371/journal.pone.0223832
8. Villasís-Keever MÁ, Márquez-González H, Zurita-Cruz JN, Miranda-Novales G, Escamilla-Núñez A. El protocolo de investigación VII. Validez y confiabilidad de las mediciones. Rev Alerg Mex. 2018;65(4):414-421. DOI: 10.29262/ram.v65i4.560
9. Díaz-García L, Medina-Vera I, García-de la Puente S, González-Garay A, Murata Ch. Estudios de exactitud diagnóstica. Acta Pediatr Mex. 2019;40(6):342-357. Disponible en: https://ojs.actapediatrica.org.mx/index.php/APM/article/download/1933/1164
10. Le-Gal G, Le-Roux PY. How to assess quality of primary research studies in the medical literature? Semin Nucl Med. 2019;49(2):115-120. DOI: 10.1053/j.semnuclmed.2018.11.007
11. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137(4):558-565. DOI: 10.5858/arpa.2012-0198-RA
12. Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. Acad Emerg Med. 2013;20(11):1194-1206. DOI: 10.1111/acem.12255
13. Ramírez-Enríquez F, Prado-Rendón J, Lachica-Valle J, Valle-Leal JG. Inmunoglobulina E total como marcador de alergia en el noroeste de México. Rev Alerg Mex. 2016;63(1):20. DOI: 10.29262/ram.v63i1.135
14. Marraccini P, Pignatti P, D Apos-Alcamo A, Salimbeni R, Consonni D. Basophil activation test application in drug hypersensitivity diagnosis: an empirical approach. Int Arch Allergy Immunol. 2018;177(2):160-166. DOI: 10.1159/000490116