Colombian Journal of Anesthesiology

OPEN

Recibido: 1 noviembre, 2023 ▸ Aceptado: 26 febrero, 2024 ▸ Online first: 2 abril, 2024

DOI: https://doi.org/10.5554/22562087.e1108

Perspectivas innovadoras sobre el valor de las pruebas diagnósticas en la práctica clínica

Innovative perspectives on the value of diagnostic tests in clinical practice

Kelly Estrada-Orozco[a-c], Juliana Cuervo[a]

[a] Programa de doctorado en Epidemiología Clínica, Departamento de Epidemiología Clínica y Bioestadística, Pontificia Universidad Javeriana. Bogotá, Colombia.
[b] Unidad de Síntesis de Evidencia y Gestión de Tecnologías, Instituto de Evaluación de Tecnologías en Salud (IETS). Bogotá, Colombia.
[c] Instituto de Investigaciones Clínicas, Facultad de Medicina, Fundación Universitaria Sanitas. Bogotá, Colombia.

Correspondencia: Cochrane, Facultad de Medicina, Universidad Nacional de Colombia, Cra. 30 calle 45, Campus Universitario. Bogotá, Colombia.

Cómo citar este artículo: Estrada-Orozco K, Cuervo J. Innovative perspectives on the value of diagnostic tests in clinical practice. Colombian Journal of Anesthesiology. 2024;52:e1108.

Email: kpestradao@unal.edu.co

Resumen

Las pruebas diagnósticas tienen características intrínsecas, como la sensibilidad, especificidad, exactitud global y las razones de verosimilitud, que definen su desempeño operacional. No es infrecuente encontrar en la literatura que se valore la prueba y se defina su utilidad clínica exclusivamente de acuerdo con estas características. En este documento se presentan varios argumentos que permiten reflexionar sobre las características que verdaderamente definen el valor de las pruebas diagnósticas en la práctica clínica. Se concluye con una perspectiva en la que el valor de cada prueba diagnóstica se establece de acuerdo con las circunstancias de uso de la misma: de quién, cuándo, dónde y en quién se use la prueba, y todas estas son características extrínsecas de una prueba diagnóstica.

Palabras clave

Pruebas diagnósticas; Diagnóstico; Valor clínico; Sensibilidad y especificidad; Práctica clínica.

Abstract

Diagnostic tests have intrinsic characteristics such as sensitivity, specificity, overall accuracy and likelihood ratios which define their operational performance. It is not uncommon to find in the literature that test value and clinical utility are defined based exclusively on those characteristics. This paper introduces several arguments aimed at prompting a reflection regarding the characteristics that define the true value of diagnostic tests in clinical practice. It concludes with the view that the value of each diagnostic test needs to be established in accordance with the circumstances in which it is used, taking into account extrinsic characteristics such as in whom it is used, when, where and by who.

Key words

Diagnosis; Diagnostic test; Clinical value; Sensitivity and specificity; Clinical practice.

Read the English version of this article on the journal website www.revcolanest.com.co

INTRODUCCIÓN

Una prueba diagnóstica es cualquier elemento capaz de modificar la probabilidad de diagnóstico de una condición. Específicamente, en la práctica clínica, las pruebas de diagnóstico son enfoques que se utilizan para identificar con alta exactitud la enfermedad en un paciente y, por lo tanto, proporcionar un tratamiento temprano y adecuado (1).

En su propósito, las pruebas pueden ser utilizadas para detección, evaluación de riesgos, diagnóstico, caracterización del pronóstico, estadificación, monitoreo o vigilancia (1). Por otra parte, en el proceso diagnóstico una prueba se puede introducir como: 1. Reemplazo (es decir, pruebas con menos carga, invasividad, costo o exactitud superior); 2. Triaje (es decir, pruebas que definen la continuación de un proceso diagnóstico y en consecuencia minimizan el uso de una prueba invasiva o costosa); 3. Adición (es decir, para mejorar la exactitud dentro del proceso diagnóstico existente), o 4. Pruebas paralelas o combinadas (muy utilizadas en clínica, estas son pruebas para iguales o distintas condiciones de salud que en el abordaje sindrómico permiten ir descartando diagnósticos diferenciales) (2).

No es infrecuente encontrar en la literatura que se valore a la prueba diagnóstica como una prueba excelente cuando esta es exacta (el valor medido es lo más cercano posible al valor real) y precisa (el valor medido es repetible y reproducible) (1,3,4) e incluso, se considere como una prueba de diagnóstico "ideal", "prueba perfecta" o "idónea" a aquella que identifique correctamente a los sujetos con y sin la enfermedad con una exactitud del 100 % (5,6). Aunque la exactitud y la precisión son características mínimas necesarias para valorar una prueba diagnóstica como una prueba ideal, no son suficientes para definir el valor y la utilidad de la prueba. Es más, su verdadero valor no depende solamente de sus características operativas intrínsecas como sensibilidad, especificidad, valores predictivos positivos y negativos o exactitud global, sino de qué tanto la prueba puede ser utilizada en un contexto específico y le aporta al usuario en su decisión clínica para proporcionar un tratamiento adecuado y oportuno que genere beneficios a los pacientes; es decir, qué tan útil es la prueba. Y las características de quién, cuándo, dónde y en quién se usa la prueba son particularidades extrínsecas de una prueba diagnóstica (7,8).

El objetivo de este documento es presentar argumentos basados en evidencia de por qué las características operativas intrínsecas —sensibilidad, especificidad y exactitud diagnóstica, entre otras, que caracterizan su validez técnica— son tan solo el punto de partida para evaluar el valor de una prueba diagnóstica. En la práctica, factores extrínsecos a la prueba y característicos del contexto clínico de su aplicación determinan su desempeño operativo y, por lo tanto, su consideración es necesaria para guiar decisiones sobre su uso y definir su verdadero valor o utilidad.

A continuación, se discuten: 1. El papel que tienen las características intrínsecas de la prueba en el proceso diagnóstico; 2. El papel de la certeza en las características intrínsecas de las pruebas en el proceso diagnóstico; 3. La variabilidad del desempeño operativo de la prueba de acuerdo con el escenario de uso y el usuario de la prueba, y 4. Otros factores que afectan el uso de las pruebas y contribuyen a definir su valor.

El papel de las características intrínsecas de las pruebas diagnósticas

La palabra intrínseco proviene del vocablo latino intrinsĕcus y se utiliza para calificar aquello que es propio de algo (9). En el escenario de las pruebas diagnósticas, las características intrínsecas corresponden a aquellas que definen su "desempeño" diagnóstico, es decir, su capacidad de clasificar correctamente a individuos con y sin la condición de interés, y entre las que sobresalen medidas estándar como la sensibilidad (Sen), especificidad (Sp), valores predictivos positivos (VPP) y negativos (VPN), exactitud global (EG), razones de verosimilitud positivas y negativas (LR+ y LR-), la razón de posibilidades diagnósticas (DOR u OR diagnóstico) y el índice de Youden (J). Otras medidas menos conocidas también han sido propuestas como resumen del "rendimiento" de la prueba (el funcionamiento de la prueba en escenarios clínicos específicos) como es el caso del número necesario para diagnosticar (NND), el número necesario para diagnosticar erróneamente (NNM), e incluso se ha propuesto un índice de utilidad clínica de un resultado positivo o negativo basado en los valores predictivos correspondientes y la sensibilidad y la especificidad, respectivamente, con umbrales que

Tabla 1. Descripción del uso del índice de Utilidad Clínica de las pruebas según A. Mitchel.

Índice de utilidad clínica (IUC)	Interpretación de su utilidad
IUC > 0,81	Excelente
0,64 ≤ IUC < 0,81	Buena
0,49 ≤ IUC < 0,64	Justa
0,36 ≤ IUC < 0,49	Poca
IUC < 0,36	Muy poca

Fuente: Autores, a partir de (11).

definen el grado en el que una prueba es "útil" en la práctica clínica (Tabla 1) (1,10-12).

Sin importar cuan elaboradas parezcan las medidas, valorar la utilidad de la prueba solamente teniendo en cuenta sus características operativas básicas, sin considerar el contexto y cómo son realmente interpretados y aplicados sus resultados, puede ser arbitrario y poco adecuado. Por ejemplo, ¿cuánto valor agregan a la decisión clínica las pruebas con mayor sensibilidad, especificidad o exactitud? O ¿cuán verdaderamente útiles son las que tienen un excelente índice de "utilidad"?

Por ejemplo, la prueba autoadministrada para VIH (autotest). Esta prueba tiene una sensibilidad de 100 % —el 100 % de las personas con infección por VIH obtienen un resultado positivo con esta prueba— y una especificidad del 99,8 % —el 99,8 % de las personas sin infección por VIH obtienen un resultado negativo—. Además, esta prueba es altamente confiable, y un estudio de factibilidad sobre su manejo por personas no profesionales mostró que más del 99,2 % de los participantes obtuvieron un resultado interpretable y más del 98,1 % interpretaron correctamente el resultado. Los resultados positivos se interpretaron correctamente en el 100 % de los casos (13).

A pesar de ser una prueba que en un contexto de alta prevalencia de infección —y, por lo tanto, de alto VPP— tendría un IUC que la clasifica como una prueba diagnóstica excelente, no hace un diagnóstico definitivo, y requiere, de acuerdo con las guías de manejo, que se realice en todos los casos de positividad una prueba confirmatoria (14). Las consecuencias de un falso positivo en este caso implicarían el inicio de terapia antirretroviral, el impacto en la salud mental de la persona y otras implicaciones sociales que obligan a la utilización de una segunda prueba para obtener el diagnóstico definitivo, estableciendo así un rol de tamizaje para la prueba autoadministrada.

La definición del rol de esta prueba no proviene simplemente de sus características operativas: es una prueba que funciona, es exacta y confiable, pero insuficiente como herramienta única de diagnóstico, porque el juicio sobre su uso implica considerar las consecuencias de un mal diagnóstico, a pesar de que sea improbable. Por otra parte, la prueba autoadministrada ofrece beneficios en cuanto al acceso al diagnóstico y la oportunidad de la atención, ya que, en caso de un resultado positivo, la prueba le permite a la persona buscar atención médica y beneficiarse del tratamiento después de someterse a una prueba de laboratorio confirmatoria. Si el tratamiento antirretroviral se inicia tempranamente, la esperanza de vida de las personas con VIH puede ser similar a la de la población general.

En este otro ejemplo, la Asociación Americana del Embarazo (AAE) recomienda la realización de pruebas caseras de embarazo mencionando que su exactitud está entre 97 % y 99 % cuando se realizan adecuadamente, e indican que son una opción rápida, de bajo costo y que garantizan la privacidad del usuario. A pesar de su alta exactitud diagnóstica, esta prueba no es suficiente cuando se trata de confirmar o descartar el embarazo. La razón es simple, un falso positivo o un falso negativo tienen consecuencias de gran impacto; por ejemplo, un falso negativo retrasaría el ingreso oportuno a programas de control prenatal, con las implicaciones de ello en la salud materna y fetal. Por esa razón, a pesar de ser una prueba con un índice de utilidad excelente, con una exactitud diagnóstica alta y valores de sensibilidad y especificidad por encima del 95 %, no sería una prueba candidata para realizar un diagnóstico definitivo.

En 2014, Josephson et al. publicaron un metaanálisis que da cuenta de la estimación combinada de la sensibilidad y especificidad tanto de la angiografía por tomografía (angioTAC o ATC) como de la angiografía por resonancia magnética (ARM) para la detección de malformaciones vasculares en pacientes con hemorragia intracerebral (15). En los estudios de ATC, la estimación combinada de la sensibilidad fue de 95 %; (intervalo de confianza [IC] 95 %: 90 a 97 %) y la especificidad fue de 99 % (IC 95 %: 95 a 100 %), y en los estudios de ARM, la estimación combinada de la sensibilidad fue 98 % (IC 95 %: 80 a 100 %) y la especificidad fue 99 % (IC 95 %: 97 a 100 %). La pregunta sobre cuál de las dos pruebas usar para tomar la decisión acerca de una conducta quirúrgica para un paciente con hemorragia intracraneana puede tener una respuesta tan simple como: usar la que se encuentre disponible o la menos costosa o la de preferencia del clínico, porque ambas tienen una alta exactitud y un excelente índice de utilidad clínica; sin embargo, otras consideraciones pueden hacer que se incline la balanza a preferir la ATC por encima de la ARM, al menos con los datos de este estudio, y tiene que ver con las consecuencias de la decisión derivadas de una mayor frecuencia de falsos negativos que pueden ocurrir al emplear una ARM (Sen IC 95 %: 80 a 100 %). Incluso, las características clínicas y antecedentes del paciente, como historia de trauma o de otras comorbilidades, pueden hacer que se defina por una u otra, lo que nuevamente indica que hay un conjunto de condiciones externas a la prueba que determinan su uso y utilidad clínica.

Se ha creído que entre más estables sean las características intrínsecas de la prueba frente a la prevalencia de la condición de interés diagnóstico (16), mejor es para la toma de decisiones clínicas, y eso ha posicionado a la alta sensibilidad y especificidad como características deseables de una prueba; lo cierto es que, por sí misma, una prueba sensible o específica seleccionada de acuerdo con su objetivo no resuelve las dificultades a las que se enfrentan sus usuarios, y contrario a lo que se piensa, las pruebas pueden ofrecer diferentes grados de información de acuerdo con la prevalencia de la condición en la población en la que se usen (17-22).

Sucede igual para otras características intrínsecas de las pruebas diagnósticas, como es el caso de las razones de verosimilitud tanto positivas como negativas. Por ejemplo, la ultrasonografía hepática y de vías biliares es considerada el patrón de oro para colecistitis aguda, en parte por las excelentes características operativas de la prueba. En atención de urgencias, el hallazgo de líquido libre pericolecístico en ultrasonografía tiene un LR + = 10,7 y un LR − =

0,8 (23); sin embargo, la probabilidad posprueba que se alcanza con su resultado positivo en un paciente con dolor abdominal agudo es tan solo del 20 %, y permanece sin cambios (~ 2 %) cuando su resultado es negativo (probabilidad preprueba de 2 %, a partir de la prevalencia de colelitiasis en población general del 5 a 10 %, y solo el 20 % de los pacientes con colelitiasis evolucionan a colecistitis) (24,25). Su verdadero valor se observa en escenarios de aplicación con probabilidades preprueba mayores al 10 %, es decir, en poblaciones clínicas seleccionadas basadas en otras pruebas diagnósticas y en la revisión de signos clínicos. Entonces, pensar que la ultrasonografía hepática y de vías biliares globalmente tiene una excelente utilidad clínica es erróneo, pues su utilidad depende de la situación en la que se aplica, es decir, es contexto-dependiente.

De acuerdo con lo expuesto, las características intrínsecas de las pruebas son necesarias, pero no suficientes para determinar su valor. No son simplemente la alta sensibilidad o especificidad o exactitud lo que determina el valor de la prueba para la toma de decisión clínica, sino que hay características del contexto o del escenario de uso que lo definen.

El papel de la certeza que se tiene en las características intrínsecas de la prueba diagnóstica

Las medidas de desempeño de las pruebas diagnósticas son valores estimados con cierto grado de incertidumbre. Para determinar las características intrínsecas de una prueba se requiere tener un comparador que tenga características operativas insuperables en el contexto en el que se aplica y para la condición de interés, esta prueba es el Gold estándar o estándar de oro. El estándar de oro se puede definir como el mejor método disponible para determinar la presencia o ausencia de la condición de interés (26), sus características no son exclusivamente operativas, ya que su uso es producto de un proceso de consenso, prueba de beneficio adicional y aceptación (2).

Aunque se reconoce la importancia de contar con una prueba de referencia con características de un estándar de oro, en la práctica diaria, la verificación de los verdaderos diagnósticos, es decir, de los sujetos que verdaderamente tienen la condición de interés mediante el estándar de oro, puede resultar poco factible, ya sea por el riesgo para el paciente, la inversión en recursos humanos e institucionales, lo poco práctico de la prueba o los conflictos éticos que representa su realización. En otros casos, como algunas enfermedades psiquiátricas —entre ellas ansiedad, depresión (27) o esquizofrenia (28,29)— el estándar de oro ni siquiera está disponible. Aunque la falta de un estándar de oro perfecto en la práctica de investigación es frecuente, no hay un consenso sobre cuál es la mejor opción para evitar introducir sesgos en la comparación de la nueva prueba frente al estándar de oro y evaluar sus características intrínsecas (30).

El término estándar de referencia o de criterio se prefiere en ausencia de un estándar de oro. La diferencia radica en que estos dos constituyen estrategias o pruebas que corresponden al mejor enfoque actual y aceptado para hacer un diagnóstico, y que permiten hacer la comparación con la prueba de interés que se quiere evaluar, sin que sea necesariamente una prueba con desempeño perfecto. En otros casos, como se explicó, el estándar de oro está disponible, pero ante los riesgos éticos o de factibilidad que limitan su uso —por ejemplo, la biopsia cerebral como estándar de oro para el diagnóstico de enfermedad de Alzheimer— se prefiere como estándar de referencia otra prueba con menor desempeño operativo (31). Entonces, la incertidumbre se hace evidente en la medida en que se están determinando las características de la prueba en estudio frente a un estándar de referencia del que existe acuerdo en que es la mejor opción disponible, pero no necesariamente es la prueba con el mejor desempeño operativo. Esto puede implicar que la nueva prueba tenga mejores características operativas para el diagnóstico que el estándar de referencia, aunque sigan siendo más bajas si se comparan contra el estándar de oro. Por ejemplo, recientemente se han propuesto biomarcadores para el cáncer de próstata como un reemplazo más exacto del antígeno prostático específico, aunque el estándar de oro es la biopsia (32).

Otros aspectos metodológicos de los estudios en los que se determinan las características intrínsecas de las pruebas diagnósticas pueden afectar la certeza en dichas medidas (33). El primer paso para evaluar el valor de una prueba médica antes de realizar estudios comparativos de impacto es la evaluación de su exactitud (34). Esta evaluación se realiza mediante estudios de corte transversal que pueden estar anidados en diseños de tipo longitudinal como cohortes, experimentos clínicos o estudios de casos y controles (34), con algunas ventajas de los primeros por presentar menor riesgo de incremento artificial de la exactitud de la prueba, secundario a valores sesgados de prevalencia (30). Sin embargo, el tipo de diseño no es la única preocupación en los estudios de exactitud diagnóstica, el riesgo de sesgos durante la selección de los participantes, la aplicación y de la interpretación de la prueba en estudio (prueba índice) y el patrón de referencia, entre otros, son otras fuentes reconocidas de incertidumbre de los resultados obtenidos (33,35,36).

Por lo tanto, para entender el valor de una prueba es necesario también entender el grado de incertidumbre que rodea a las medidas sobre su capacidad discriminativa y su confiabilidad, y la posibilidad de que estén sesgadas y subestimen o sobreestimen la exactitud real.

Comportamiento de las características intrínsecas de acuerdo con el escenario de uso y el usuario de la prueba

La confiabilidad de la prueba se refiere a la variación entre las mediciones hechas por la prueba sobre una unidad de análisis, que se explica por errores de medición (37), ya sea por problemas de repetibilidad o de reproducibilidad. La repetibilidad en teoría de la medición se refiere a la variación

en las mediciones realizadas en diferentes momentos sobre la misma unidad de análisis en condiciones idénticas, que, en caso de existir, son atribuibles a errores del proceso de medición. Para comprobar que existe repetibilidad se requiere que las mediciones se hagan usando un mismo instrumento o método, el mismo observador o evaluador, y que las mediciones se realicen durante un periodo en el que no se espera que ocurra variación en el registro de interés (37).

La reproducibilidad, por otra parte, se refiere a la variación de las mediciones realizadas sobre una unidad de análisis en condiciones que no son idénticas, ya sea porque se espera que existan cambios en el tiempo sobre la unidad de análisis objeto de medición o por el uso de diferentes métodos, instrumentos o evaluadores (37).

Una prueba diagnóstica puede tener excelentes características intrínsecas, incluyendo una buena reproducibilidad y repetibilidad, sin embargo, su verdadera utilidad dependerá del uso que se le dé a la prueba. Por ejemplo, las pruebas serológicas detectan anticuerpos o inmunoglobulinas que se producen como respuesta inmune humana a la infección. Cuando los anticuerpos de inmunoglobulina M (IgM) están presentes, pueden indicar una infección activa o reciente; mientras que los anticuerpos de inmunoglobulina G (IgG) aparecen más tarde en el proceso de infección y a menudo pueden indicar una infección pasada, pero no excluye a los pacientes infectados recientemente (38).

Las pruebas serológicas pueden ser importantes para la detección temprana de infecciones. Estas pruebas son fáciles de operar y permiten un cribado rápido de anticuerpos en un plazo de 10 a 15 minutos y, debido a su bajo costo y fácil y rápido procesamiento, en algunos escenarios se utilizan como una herramienta de detección para la población general (39).

Para detección de la infección por SARS-CoV-2 se han desarrollado pruebas de anticuerpos para detectar solo IgG, tanto IgG como IgM, o anticuerpos totales; sin embargo, las características operativas de estas pruebas varían de manera significativa de acuerdo con el momento clínico en el que se realizan, así como por las características de los pacientes en los que se aplican. Las pruebas de anticuerpos realizadas una semana después de los primeros síntomas solo detectan el 30 % de las personas que tienen COVID-19, esta cifra aumenta a 70 % en la segunda semana y a más del 90 % en la tercera semana (40). Por otra parte, en pacientes asintomáticos la sensibilidad combinada de IgM es 28,6 % (IC 95 %: 23,8-33,7 %); en pacientes sintomáticos, entre 8-11 días o menos desde el inicio de los síntomas la sensibilidad combinada para IgM es de 33 % (IC 95 %: 23-43 %), y en pacientes sintomáticos con más de 11 días desde el inicio de los síntomas la sensibilidad para IgM es 66 % (IC 95 %: 61-70 %) (39).

Como se observa en el ejemplo, la sensibilidad de la prueba varía según las características del sujeto (el estar asintomático o sintomático) y al tiempo transcurrido desde la exposición o el inicio de síntomas. Nuevamente, es claro que las características intrínsecas de la prueba no definen de forma absoluta su utilidad clínica. Para este caso particular, su desempeño es variable según el momento de la historia de la enfermedad en el que se utilice, lo cual indica que es necesario saber cuándo usar la prueba diagnóstica, reconocer su papel en el proceso diagnóstico y comprender cómo funciona y por qué se usa la prueba, esto indiscutiblemente requiere unos mínimos de experiencia del usuario de la misma.

Otras implicaciones del uso de las pruebas diagnósticas

Pensar en las implicaciones automáticamente ubica de nuevo en las características extrínsecas de la prueba. Aunque se ha avanzado en considerar las consecuencias del uso de las pruebas a partir de los resultados falsos positivos y falsos negativos como tratar de más o dejar de tratar a los pacientes, las implicaciones del uso de la prueba requieren reflexiones más allá de lo que se deriva de estas características intrínsecas y exigen considerar también los recursos económicos y humanos necesarios para aplicar la prueba, así como juzgar el balance riesgo-beneficio de sus resultados desde una perspectiva social y ética.

Tan relevantes son estos aspectos que, en algunos contextos, la prueba de más valor no es la más exacta, sino aquella disponible para tomar de forma oportuna una decisión que permita modificar el curso clínico de un paciente cuando no hay otras opciones disponibles. Esto puede ser incluso una anamnesis clara, bien dirigida y semiológicamente rica en detalles.

En condiciones de recursos muy limitados o personal poco capacitado, pruebas muy exactas, pero de implementación o interpretación complejas, pueden ser de poca utilidad y valor, mientras que pruebas con buena, aunque quizá inferior exactitud, pero económicas, rápidas y fácilmente realizables e interpretables con un mínimo de entrenamiento pueden resultar de gran utilidad y valor para una población.

Por otra parte, pudiera no ser ético diagnosticar a pacientes con condiciones para las que no se tiene cómo intervenir para curar o modificar el curso clínico. Realizar una prueba en estas condiciones tiene el potencial de violar cualquiera de los cuatro principios éticos, a saber, beneficencia, no maleficencia, justicia y autonomía. Un ejemplo de prueba diagnóstica en la que potencialmente riñen sus excelentes características intrínsecas (100 % de sensibilidad y 98,9 % de especificidad) (41) y su utilidad y valor al considerar aspectos extrínsecos, son las pruebas genéticas en enfermedad de Alzheimer (EA).

La EA de inicio tardío es la forma más común de esta patología y suele ser esporádica. Sin embargo, se han identificado algunos alelos que aumentan el riesgo de desarrollar EA. APOE ε4 es un factor de riesgo bien establecido para la EA y se asocia con un riesgo cuatro veces mayor de desarrollar la enfermedad (42,43). Si bien las pruebas genéticas pueden identificar fácilmente la presencia o ausencia de estos genes de susceptibilidad, esto tiene poco beneficio clínico o diagnóstico, puesto que no hay tratamiento modificador del riesgo, y adicionalmente, permanece la incertidumbre diagnóstica dado que un paciente puede

ser portador del alelo APOE ε4 y no desarrollar EA, o puede desarrollar EA sin el alelo APOE ε4 (42). Entonces, ¿qué utilidad clínica tiene la prueba si no se puede proporcionar un tratamiento temprano y adecuado? Además, el conocimiento del estado de portador puede generar una enorme carga emocional ante la incertidumbre y la imposibilidad actual de intervenciones efectivas.

Otro ejemplo en el que la utilidad de la prueba se define por sus características extrínsecas, a pesar de sus excelentes características intrínsecas, es el diagnóstico de COVID-19. La prueba de oro para el diagnóstico es RT-PCR (en inglés Reverse transcription polymerase chain reaction) que alcanza una sensibilidad del 85,7 % (IC 95 %: 81,5-89,1 %) en pacientes hospitalizados; 95,5 % (IC 95 %: 92,2-97,5 %) en pacientes ambulatorios y 89,9 % (IC 95 %: 88,2-92,1 %) en todos los pacientes (44). Sin embargo, la disponibilidad de la prueba en algunas regiones es escasa y los tiempos para entrega de resultados son prolongados; además, la susceptibilidad de la prueba a fallas en la toma, transporte y procesamiento y los costos de la misma, hacen que no sea la prueba de mayor utilidad clínica; por el contrario, las pruebas rápidas antigénicas (Ag-RDT) (sensibilidad 84 a 97 % y especificidad 97 a 100 % comparada con RT-PCR [45]) se realizan muy rápidamente y son más fáciles de usar e interpretar. Las Ag-RDT brindan un resultado en menos de 30 minutos, lo que puede contribuir al diagnóstico, al rastreo y estudio de contactos y, por lo tanto, a frenar la transmisión de SARS-CoV-2 en una comunidad (46).

CONCLUSIONES

A partir de lo que se argumenta en este documento es posible concluir que en la práctica clínica y de salud pública, la utilidad y el valor de una prueba no los definen exclusivamente sus características intrínsecas, sino que el valor de cada prueba diagnóstica se establece de acuerdo con las circunstancias de uso de la misma: de quién, cuándo, dónde y en quién se use la prueba, y estas son características extrínsecas de una prueba diagnóstica. Es necesario, por lo tanto, un ejercicio reflexivo y sistemático que permita tomar la decisión sobre el uso o introducción de una prueba basados no solo en cuáles son sus características intrínsecas y la certeza que se tiene de su desempeño, sino también, y especialmente, valorando la prueba de acuerdo con las circunstancias que motiven su uso y el contexto en el que este se da, particularmente, las probabilidades preprueba, las consecuencias de dejar pasar un diagnóstico o sobrediagnosticar, los riesgos de usar la prueba, la factibilidad de su aplicación correcta, su aceptabilidad e interpretabilidad, la disponibilidad, los costos y demás recursos necesarios para su utilización, y las consecuencias de su uso desde una perspectiva ética. En conclusión, desde el punto de vista de los autores de este artículo, no hay una única prueba diagnóstica ideal o mejor para una condición; existen pruebas que aportan valor a la decisión clínica de acuerdo con cada escenario de uso y del contexto.

Conflictos de interés

KEO es miembro del grupo GRADE y del GRADE Diagnosis Group. No hay otros conflictos que declarar por los autores.

Financiación

Ninguna declarada por las autoras.

REFERENCIAS

1. Bolboacă SD. Medical Diagnostic Tests: A review of test anatomy, phases, and statistical treatment of data. Comput Math Methods Med. 2019; 1891569. doi: https://doi.org/10.1155/2019/1891569
2. Schünemann HJ, Mustafá RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-41. doi: https://doi.org/10.1016/j.jclinepi.2019.12.020
3. Šimundić AM. Measures of diagnostic accuracy: Basic definitions. EJIFCC. 2009;19(4):203-11.
4. Shreffler JHM. Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [citado: Enero 16 de 2024]. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK557491/
5. Wong HB. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. Proceed Singapore Healthc. 2001;20(4):316-8. doi: https://doi.org/10.1177/201010581102000411
6. Pluddemann A BA, O'Sullivan J. Spectrum bias: Sackett Catalogue Of Bias [internet]. 2019 [citado: Enero 16 de 2024]. Disponible en: https://catalogofbias.org/biases/spectrum-bias/
7. Buehler AM, Ascef BdO, Oliveira HAd, Ferri CP, Fernandes JG. Rational use of diagnostic tests for clinical decision making. Revista da Associação Médica Brasileira. 2019;65. doi: https://doi.org/10.1590/1806-9282.65.3.452
8. Schünemann HJ, Mustafá RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82. doi: https://doi.org/10.1016/j.jclinepi.2019.02.003
9. Definición de intrínseco [internet]. 2024 [citado: 16 enero de 2024]. Disponible en: https://definicion.de/intrinseco/
10. Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. Epidemiology. 2013;24(1):170. doi: https://doi.org/10.1097/EDE.0b013e31827825f2
11. Mitchell AJ. The clinical significance of subjective memory complaints in the diagnosis of mild cognitive impairment and dementia: a meta-analysis. Int J Geriatr Psychiatry. 2008;23(11):1191-202. doi: https://doi.org/10.1002/gps.2053
12. Mitchell AJ. Sensitivity × PPV is a recognized test called the clinical utility index (CUI+). Eur J Epidemiol. 2011;26(3):251-2. doi: https://doi.org/10.1007/s10654-011-9561-x
13. Santé. Autotest VIH France [internet]. 2024 [citado: Enero 16 de 2024]. Disponible en: https://www.autotest-sante.com/en/autotest-VIH-par-AAZ-139.html
14. Ministerio de Salud y de Protección Social. Guía de práctica clínica basada en la evidencia para la atención de la infección por VIH/SIDA en personas adultas, gestantes y adolescentes. Colombia: Minsalud; 2022.

15. Josephson CB, White PM, Krishan A, Al-Shahi Salman R. Computed tomography angiography or magnetic resonance angiography for detection of intracranial vascular malformations in patients with intracerebral haemorrhage. Cochrane Database Syst Rev. 2014;2014(9):Cd009372. doi: https://doi.org/10.1002/14651858.CD009372.pub2

16. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract. 2006;12(2):132-9. doi: https://doi.org/10.1111/j.1365-2753.2005.00598.x

17. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013;185(11):E537-44. doi: https://doi.org/10.1503/cmaj.121286

18. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598-602. doi: https://doi.org/10.7326/0003-4819-137-7-200210010-00011

19. Feinstein AR. Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health. 2002;56(5):330-2. doi: https://doi.org/10.1136/jech.56.5.330

20. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16(9):981-91. doi: https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<981::aid-sim510>3.0.co;2-n.

21. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12. doi: https://doi.org/10.1016/j.jclinepi.2008.04.007

22. Hultcrantz M, Mustafá RA, Leeflang MMG, Lavergne V, Estrada-Orozco K, Ansari MT, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. J Clin Epidemiol. 2020;117:138-48. doi: https://doi.org/10.1016/j.jclinepi.2019.05.002

23. Jang TB, Ruggeri W, Kaji AH. The predictive value of specific emergency sonographic signs for cholecystitis. J Med Ultras. 2013;21(1):29-31. doi: https://doi.org/10.1016/j.jmu.2013.01.006

24. Zarate AJ. ÁM, King, I, Torrealba. A. Colecistitis aguda. Universidad Finis Terrae: Escuela de Medicina 2016;7.

25. Halpin V. Acute cholecystitis. BMJ Clin Evid. 2014;2014.

26. Gelaye B, Tadesse MG, Williams MA, Fann JR, Vander Stoep A, Andrew Zhou X-H. Assessing validity of a depression screening instrument in the absence of a gold standard. Ann Epidemiol. 2014;24(7):527-31. doi: https://doi.org/10.1016/j.annepidem.2014.04.009

27. Davison TE, McCabe MP, Mellor D. An examination of the "gold standard" diagnosis of major depression in aged-care settings. Am J Geriatr Psychiatry. 2009;17(5):359-67. doi: https://doi.org/10.1097/JGP.0b013e318190b901

28. Wood SJ, Yung AR. Diagnostic markers for schizophrenia: do we actually know what we're looking for? World Psychiatry. 2011;10(1):33-4. doi: https://doi.org/10.1002/j.2051-5545.2011.tb00006.x

29. van Os J, Tamminga C. Deconstructing psychosis. Schizophr Bull. 2007;33(4):861-2. doi: https://doi.org/10.1093/schbul/sbm066

30. Estrada-Orozco K. Diseño de una prueba para diagnóstico de trastorno cognitivo y validación en una cohorte de sujetos mayores de 50 años en Colombia en el 2016-2017. Bogotá, D.C: Universidad Nacional de Colombia; 2018.

31. Pietrzak K, Czarnecka K, Mikiciuk-Olasik E, Szymanski P. New perspectives of alzheimer disease diagnosis - the most popular and future methods. Med Chem. 2018;14(1):34-43. doi: https://doi.org/10.2174/1573406413666171002120847

32. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ (Clinical research ed). 2006;332(7549):1089-92. doi: https://doi.org/10.1136/bmj.332.7549.1089

33. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174(4):469-76. doi: https://doi.org/10.1503/cmaj.050090

34. Mathes T, Pieper D. An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews. Systematic Reviews. 2019;8(1):226. doi: https://doi.org/10.1186/s13643-019-1131-4.

35. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137(4):558-65. doi: https://doi.org/10.5858/arpa.2012-0198-RA

36. Whiting P, Rutjes A, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36. doi: https://doi.org/10.7326/0003-4819-155-8-201110180-00009

37. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultr Obstet Gynecol. 2008;31(4):466-75. doi: https://doi.org/10.1002/uog.5256

38. Mahajan A, Manchikanti L. Value and validity of coronavirus antibody testing. Pain Physician. 2020;23(4s):S381-s90. doi: https://doi.org/10.36076/ppj.2020/23/S381

39. Mercado M, Malagón-Rojas J, Delgado G, Rubio VV, Muñoz Galindo L, Parra Barrera EL, et al. Evaluation of nine serological rapid tests for the detection of SARS-CoV-2. Rev Panam Salud Pública. 2020;44:e149. doi: https://doi.org/10.26633/RPSP.2020.149

40. Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Spijker R, Taylor-Phillips S, et al. Antibody tests for identification of current and past infection with SARS-CoV-2. Cochrane Database of Systematic Reviews. 2020(6). doi: https://doi.org/10.1002/14651858.CD013652.pub2

41. Veiga S, Rodríguez-Martín A, García-Ribas G, Arribas I, Menacho-Román M, Calero M. Validation of a novel and accurate ApoE4 assay for automated chemistry analyzers. Scientific Reports. 2020;10(1):2138. doi: https://doi.org/10.1038/s41598-020-58841-7

42. Atkins ER, Panegyres PK. The clinical utility of gene testing for Alzheimer's disease. Neurol Int. 2011;3(1):e1-e. doi: https://doi.org/10.4081/ni.2011.e1

43. Bertram L, Tanzi RE. Genome-wide association studies in Alzheimer's disease. Hum Mol Genet. 2009;18(R2):R137-45. doi: https://doi.org/10.1093/hmg/ddp406

44. Kortela E, Kirjavainen V, Ahava MJ, Jokiranta ST, But A, Lindahl A, et al. Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients. PloS One. 2021;16(5):e0251661-e. doi: https://doi.org/10.1371/journal.pone.0251661

45. Peeling RW, Olliaro PL, Boeras DI, Fongwen N. Scaling up COVID-19 rapid antigen tests: promises and challenges. The Lancet Infectious Diseases. 2021;21(9):E290-5. doi: https://doi.org/10.1016/S1473-3099(21)00048-7

46. World Health Organization. WHO provides one million antigen-detecting rapid diagnostic test kits to accelerate COVID-19 testing in Indonesia. World Health Organization [internet]. 2021 [citado: 16 Enero 2024]. Disponible en: https://www.who.int/indonesia/news/detail/17-03-2021-who-provides-one-million-antigen-detecting-rapid-diagnostic-test-kits-to-accelerate-covid-19-testing-in-indonesia.