

Validez de estudios de tests diagnósticos

Lorena Valenzuela D^{1,2}, Lorena Cifuentes A^{1,3}.

Validity of diagnostic tests

En un artículo anterior de esta serie, “El proceso diagnóstico”¹, se explicó cómo, en este proceso, partimos desde una probabilidad diagnóstica que denominamos probabilidad *pre test* cuando nos enfrentamos a un paciente. El objetivo del clínico será entonces modificar esta probabilidad *pre test* de modo de idealmente “cruzar” uno de los dos umbrales de decisión, el de tratamiento o el de diagnóstico. En el primer caso, decidirá el inicio de tratamiento y en el otro extremo, más allá del umbral diagnóstico, descartará el diagnóstico. En ambos casos, el paciente no será sometido a otros *tests* adicionales. Esta modificación desde la probabilidad *pre a post test*, tan crucial en nuestro quehacer clínico diario, estará directamente determinada por el *test* que se aplique. Este *test* corresponde a cualquier elemento informativo que se agregue, tal como se explicó en el artículo anterior. Esta nueva información o *test*, podrá corresponder a elementos de la historia, examen físico o algún examen de laboratorio. Cualquiera sea su naturaleza, será muy importante para el clínico utilizar un *test* confiable y que realmente ayude a modificar las probabilidades, de modo de facilitarle la toma de decisiones. Enfrentados, entonces, a un determinado *test*, cabe preguntarse si existe alguna forma de evaluar si éste nos será de real utilidad y en qué magnitud. Esta respuesta habrá que buscarla en estudios que evalúen las propiedades de un *test* diagnóstico.

Dentro del análisis crítico de estos estudios que evalúan *tests* diagnósticos, el primer paso será entonces evaluar la validez de los resultados, para posteriormente analizar los resultados en sí y su aplicabilidad. El objetivo del presente artículo es describir y analizar los principios relevantes para evaluar eficientemente la validez de los resultados de un artículo sobre un *test* diagnóstico.

En próximos artículos se revisará cómo interpretar y aplicar los resultados de estos estudios.

ENFRENTADOS AL PROBLEMA

Los médicos nos vemos comúnmente enfrentados al dilema de cuándo solicitar exámenes y cuando éstos son solicitados, de cómo interpretarlos. Con el desarrollo sostenido de la tecnología médica y la constante implementación de nuevos exámenes, los clínicos deben desarrollar cada vez más la habilidad de encontrar y evaluar estudios que intentan introducir estos exámenes o *test* diagnósticos. La Medicina Basada en Evidencia nos entrega herramientas de gran utilidad para este proceso de búsqueda, selección, análisis y posterior aplicación de los resultados encontrados².

Utilizaremos una situación clínica para ejemplificar este proceso. Durante la reunión de ingreso al Servicio de Pediatría, el residente de turno presenta 2 casos similares, de lactantes de 1 y 2 meses, con cuadro febril sin foco clínico evidente, razón por la cual fueron hospitalizados. Ambos tenían recuento de leucocitos, examen de orina y líquido cefalorraquídeo normal. Esta situación se repite con frecuencia en lactantes menores de 3 meses, pero estando en plena campaña de invierno con alta prevalencia de influenza, Ud. se pregunta si podrían implementar algún examen que permita el diagnóstico precoz de influenza en niños, con el fin de evitar hospitalizaciones innecesarias y, por otro lado, aislar oportunamente a aquellos pacientes hospitalizados que presenten influenza.

Frente a esta situación, Ud. se plantea la siguiente pregunta de diagnóstico, estructurada en 4 partes, para poder realizar posteriormente una búsqueda eficiente³:

¹Unidad de Medicina Basada en Evidencia, Escuela de Medicina, Pontificia Universidad Católica de Chile.

²Departamento de Medicina Familiar, Pontificia Universidad Católica de Chile.

³Departamento de Pediatría, Pontificia Universidad Católica de Chile, Santiago de Chile.

Correspondencia a: Dra. Lorena Valenzuela, e-mail: lvalenzu@med.puc.cl

Paciente: Niños con sospecha de influenza
 Intervención: *Test* rápido de influenza
 Comparación: Estándar de oro (*Gold standard*)
 Outcome: Diagnóstico de influenza

En relación a esta pregunta, Ud. realiza una búsqueda⁴ en Clinical Queries de PubMed, seleccionando "*diagnosis*" y "*narrow, specific*" con los términos MeSH "Influenza, Human" AND "Immunoassay". Con esta búsqueda, Ud. obtiene 50 referencias de las cuales selecciona el artículo: "*Bedside diagnosis of Influenzavirus Infections in Hospitalized Children*"⁵, ya que le parece que podría entregar la información más relevante y adecuada para responder su pregunta. Resumimos el estudio en el siguiente cuadro:

Población (P)	Inclusión (233 pacientes): <ol style="list-style-type: none"> Menores de 19 años hospitalizados con síntomas respiratorios (rinorrea, odinofagia, tos, taquipnea o apnea) Menores de 3 años hospitalizados con fiebre. Exclusión: <ol style="list-style-type: none"> Contacto con el paciente posterior a las 24 horas de ingreso. Rechazo a participar Imposibilidad de contactar a los padres
Intervención (I)	<i>QuickVue Influenza test</i> . (Test rápido para influenza, tomado al lado de la cama del enfermo).
Comparación (C) (<i>Gold Standard</i>)	<ol style="list-style-type: none"> Cultivo positivo o Dos PCRs consecutivas positivas para influenza A o B.
Outcome (O)	Diagnóstico de influenza.

VALIDEZ DE LOS RESULTADOS DE UN ESTUDIO DE DIAGNÓSTICO

Al igual que en el análisis crítico de estudios sobre intervenciones terapéuticas⁶, una vez que determinamos que un estudio sobre un *test* diagnóstico es potencialmente relevante para nuestro quehacer clínico, debemos decidir si los resultados del estudio son válidos. Esta validez se refiere a que la metodología utilizada en el estudio nos dé la suficiente confianza de que se evitaron los sesgos propios de estos diseños.

Inicialmente los estudios que evalúan las propiedades de un *test* diagnóstico se realizan con un diseño conocido como "corte transversal" en que todos los pacientes son sometidos al *test* y al *Gold Standard* como es el caso del artículo seleccionado. Este tipo de diseño es muy apropiado y frecuentemente utilizado para evaluar la exactitud del *test* para diagnosticar la enfermedad o condición de interés⁷.

Sin embargo, debemos considerar que el fin último de aplicar un examen diagnóstico a nuestros pacientes es que obtengan un beneficio en comparación a no ser sometidos al examen, es decir, debemos considerar el impacto del *test* sobre aquellos *outcomes* relevantes para el paciente. La manera más rigurosa de evaluar este impacto o beneficio del *test* es mediante un ensayo clínico en que los pacientes son distribuidos aleatoriamente a recibir el *test*, a no recibirlo o recibir métodos diagnósticos alternativos. De esta forma, se pueden medir los *outcomes* derivados de la conducta clínica adoptada en los distintos grupos según el resultado del *test*: mortalidad, morbilidad, costo, satisfacción, etc. Es así como los ensayos clínicos randomizados (ECR) son considerados la mejor evidencia para evaluar un *test* diagnóstico. A pesar de esto, no todos los *test* son ni deben ser sometidos a un complejo y costoso ECR. Este es el caso de *tests* simples que ayudan al diagnóstico de patologías con tratamiento claramente beneficioso, como por ejemplo una radiografía de tórax en una neumonía. Este examen claramente ayuda en el diagnóstico de neumonía y la conducta clínica que deriva de su resultado, el uso de antibióticos, sin duda alguna beneficiará al paciente. Distinto es el caso de exámenes más invasivos y de alto costo en que es perentorio conocer todas o la mayor cantidad de consecuencias sobre el paciente derivadas de la aplicación del *test*. Otra situación en que es crucial realizar un ECR, es la evaluación de un nuevo *test* de *screening*. Previa a la implementación de un examen a un gran número de pacientes aparentemente sanos, se debe tener certeza y por lo tanto, haber demostrado, que los beneficios sobre el bienestar de los pacientes sobrepasan los potenciales riesgos de un *screening* (costo, falsos positivos y sus consecuencias, etc.)^{7,8}.

Volviendo a la evaluación de un estudio acerca de la exactitud de un *test* con un diseño de corte transversal, situación a la que los clínicos se ven enfrentados con alta frecuencia, se deben respon-

der las siguientes preguntas para determinar la validez de sus resultados:

Crterios de validez en un estudio de un test diagnóstico

¿Se incluyó un espectro apropiado de pacientes similares a los cuales se aplicará el *test* en la práctica clínica?

¿Hubo comparación ciega con un estándar de oro (*gold standard*) independiente y adecuado?

¿Se realizó el *gold standard* independientemente del resultado del *test*?

A continuación describimos y explicamos la importancia de cada una de estas preguntas:

a) ¿Se incluyó un espectro apropiado de pacientes similares a los cuales se aplicará el *test* en la práctica clínica? Si el *test* es aplicado a pacientes con una enfermedad en estadio tardío (muy enfermos) y en pacientes sanos, los resultados no serán reales ni aplicables, ya que es muy probable que cualquier *test* pesquise a los “muy enfermos” y no pesquise a los sanos, dando una falsa impresión de las propiedades del *test*. Una situación que ejemplifica muy bien este punto y que es frecuentemente citada, es lo que ocurrió con el antígeno carcinoembrionario (ACE) en pacientes con cáncer colorectal⁹. La evaluación incluyó a pacientes en etapa avanzada de cáncer colorectal y a pacientes con otras patologías. En 35 de los 36 pacientes con cáncer avanzado, los niveles de ACE resultaron elevados, no así en los otros pacientes. Esto llevó a plantear la posible utilidad del ACE como herramienta diagnóstica para el cáncer colorectal. Sin embargo, estudios posteriores en pacientes en etapa más temprana de cáncer colorectal y por lo tanto, “menos enfermos”, y en pacientes con otro cáncer o patología gastrointestinal, es decir, condiciones potencialmente confundentes con un cáncer colorectal, cuestionaron fuertemente la exactitud y utilidad diagnóstica del ACE. Consecuentemente, el ACE fue abandonado como examen diagnóstico, probándose su utilidad sólo como un elemento en el seguimiento de pacientes con cáncer colorectal conocido.

En resumen, un espectro apropiado de pacientes, debe incluir a pacientes que comparten síntomas, pero que portan distintas patologías y excluir a los extremos, es decir, a los sanos y a los muy enfermos. Esto hará que la población de estudio se asemeje a la población de “incertidumbre diagnóstica” de la práctica

cotidiana en que se utilizará posteriormente el *test*. Esta población de “incertidumbre diagnóstica” es donde un nuevo examen tendrá su mayor utilidad ya que permitirá discriminar entre enfermedades distintas, pero parecidas en su presentación clínica¹⁰.

En nuestro ejemplo, los 233 pacientes ingresados al estudio para el *test* rápido de influenza presentaban un cuadro de enfermedad respiratoria alta caracterizado por rinorrea, odinofagia, tos, taquipnea o apnea, síntomas que pueden corresponder a una gama amplia de patologías entre las que se incluye la influenza. La población seleccionada corresponde, por lo tanto, a un espectro de pacientes con suficiente incertidumbre diagnóstica como para determinar el poder diagnóstico del *test*.

b) ¿Hubo comparación ciega con un *gold standard* independiente y adecuado? En realidad debemos responder a 3 preguntas en este punto.

- 1) ¿El *gold standard* utilizado es el más adecuado para certificar el diagnóstico o condición? Para evaluar la precisión de un *test* diagnóstico es necesario compararlo con la “verdad” (es decir, lo más cercano a la verdad disponible). Esta “verdad” suele ser el patrón de oro o *gold standard*, que debe estar aceptado como tal por la comunidad médica para considerarlo adecuado. Si el *gold standard* utilizado no parece el más apropiado, es improbable que el artículo entregue resultados válidos para nuestros propósitos. En ocasiones no existe un único *gold standard*, en estos casos, la combinación de exámenes como *gold standard* permite aumentar la precisión del diagnóstico.
- 2) Además de ser adecuado, el *gold standard* debe ser independiente del *test*. Esta independencia se refiere a que el *test* no sea parte del *gold standard* utilizado, es decir, que este *gold standard* no “contenga” al *test* como criterio, lo cual llevaría a sobredimensionar el poder diagnóstico del *test*. Por ejemplo, un estudio evaluó la utilidad de la medición de amilasa sérica y urinaria en el diagnóstico de pancreatitis aguda, considerando como *gold standard* más adecuado un listado de criterios dentro de los cuales se incluían la amilasa sérica y urinaria, lo que produce un falso aumento del poder discriminatorio del *test*.
- 3) Si el *gold standard* parece el más adecuado y es independiente, la siguiente pregunta es si los resultados del *test* y del *gold standard* fueron evaluados en forma ciega entre sí, es decir, que el investigador que interpreta el *test* no está en conocimiento del resultado ni había aplicado el

gold standard. Esto permite disminuir el sesgo de “sobreinterpretación” cuando el test diagnóstico es positivo y de “subinterpretación” cuando el test diagnóstico es negativo. Cuanto más probable sea que los resultados de un *test* puedan influir en la interpretación del otro, más importante es la interpretación ciega de ambos.

En nuestro artículo para el escenario clínico de influenza, la infección fue definida como la presencia de un cultivo positivo o 2 PCRs consecutivas positivas para influenza A o B, actualmente aceptadas por la comunidad médica como diagnóstico de certeza. El *gold standard* utilizado no incluye de modo alguno al *test* rápido en estudio, por lo que podemos confirmar su independencia. Por otro lado, a los pacientes ingresados a este estudio se les tomaron 2 muestras nasales, en dos tórulas separadas. La tórula para el *test* rápido era analizada por un investigador al lado de la cama del enfermo y la tórula para el cultivo y PCR por un tecnólogo del laboratorio que no sabía el resultado del *test* rápido, existiendo por lo tanto, una comparación ciega.

c) ¿Se realizó el *gold standard* sin considerar el resultado del *test*? Es importante que a todos los pacientes, sin importar el resultado del *test* en evaluación, se les aplique el *gold standard*. Puede ocurrir que las propiedades del *test* diagnóstico se alteren si el resultado de éste influye sobre la decisión de qué pacientes serán sometidos a continuación al *gold standard*. En algunos estudios se aplica el *gold standard* sólo a los pacientes que presentan un

resultado positivo en el *test*, especialmente si el *gold standard* es invasivo o presenta riesgos importantes; esto se denomina sesgo de verificación. Este sesgo hará inevitablemente aparecer con mayor poder diagnóstico al *test* al seleccionar pacientes con mayor probabilidad de tener un *gold standard* positivo.

Volviendo a nuestro estudio, a los 233 pacientes enrolados se les aplicaron ambos exámenes. A todos se les tomaron dos muestras nasales y se les aplicó tanto el *test* rápido como el *gold standard*.

En resumen, en nuestro escenario clínico, el artículo analizado cumple con todos los criterios de validez, por lo que podemos asumir que los resultados que nos presentarán los autores serán “creíbles”, al reducirse la posibilidad de sesgo.

CONCLUSIONES

Al realizar el análisis crítico de estudios sobre *tests* diagnósticos, se debe evaluar la validez de los resultados. Para tales efectos se deben considerar las tres preguntas anteriormente analizadas, referentes al espectro de pacientes estudiados, a la comparación ciega entre *test* y *gold standard* y a la independencia de este último, y a la presencia o ausencia de sesgo de verificación. Si se cumplen los criterios mencionados, se puede concluir que los resultados del estudio probablemente representen una estimación no sesgada de las características del *test*.

En próximos artículos de esta serie se analizará la interpretación de los resultados y su aplicabilidad en la clínica diaria.

REFERENCIAS

1. CAPURRO ND, RADA GG. El proceso diagnóstico. *Rev Méd Chil* 2007; 135: 534-8.
2. JAESCHKE R, GUYATT G, LIJMER. Diagnostic tests. En: Guyatt G, Rennie D, ed. *Users' Guides to the Medical Literature. A Manual of Evidence-based Clinical Practice*, AMA Press 2002: 121-6.
3. SOTO M, RADA G. Formulación de Preguntas en Medicina Basada en la Evidencia. *Rev. Méd. Chile* 2003; 131: 1202-3.
4. RADA G, ANDRADE M, LEYTON V, PACHECO C, RAMOS E. Búsqueda de información en medicina basada en evidencia. *Rev Méd Chile* 2004; 132: 253-9.
5. POEHLING K, GRIFFIN M, DITTUS R, TANGT Y, HOLLAND K, LI H, EDWARDS KM. Bedside diagnosis of influenza virus infections in hospitalized children. *Pediatrics* 2002; 110: 83-8.
6. PANTOJA T, LETELIER LM, NEUMANN I. El análisis crítico de la información publicada en la literatura médica. *Rev Méd Chil* 2004; 132: 513-5.
7. NEWMAN T, BROWNER W, CUMMINGS S, HULLEY S. Designing studies of medical tests. En: Hulley S, Cummings S, Browner W, Grady D, Newman T, (eds). *Designing Clinical Research*. Lippincott Williams & Wilkins 2007: 183-205.
8. GUYATT G, SACKETT D, HAYNES B. Evaluating diagnostic tests. En: Haynes B, Sackett D, Guyatt G, Tugwell D. *Clinical Epidemiology. How to do clinical practice research*. Lippincott Williams & Wilkins 2006: 273-322.
9. THOMSON DM, KRUPY J, FREEDMAN So, GOLD P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci USA*. 1969; 64: 161-7.
10. Is this evidence about a diagnostic test valid? En: Sackett D, Richardson Ws, Rosenberg W, Haynes B, ed. *Evidence-based Medicine: How to Practice & Teach EBM*. Churchill Livingstone 1998; 81-4.