Ronda clínica y epidemiológica

OSCAR OSÍO URIBE[1], JOHN JAIRO ZULETA TOBÓN[2]

¿PARA QUÉ SIRVEN Y CÓMO SE LEEN LOS ARTÍCULOS DE PRUEBAS DIAGNÓSTICAS?

"**LAS IMPRESIONES MENTALES** son de cuatro tipos: cosas que son lo que parecen ser; o no lo son y no parecen serlo; o lo son y no parecen serlo; o no lo son, aunque parecen serlo. Es misión del hombre sabio tomar la decisión correcta en todos los casos". Epicteto I-II d. C.

"El criterio más importante para determinar la utilidad de una prueba diagnóstica es saber si le proporcionará al clínico información más allá de la que está disponible por otros medios y si este nuevo conocimiento llevará a cambios que beneficien finalmente al paciente".

Un tratamiento médico efectivo usualmente depende de la precisión con la cual el médico tratante haga el diagnóstico de la enfermedad o condición que afecta a su paciente. El diagnóstico médico (dia: a través de, gnosis: conocimiento) es un proceso imperfecto que resulta en una probabilidad, más que en una certeza absoluta de haber logrado establecer la verdad acerca de la condición que padece el paciente. La mayoría de las veces el clínico aplica las pruebas diagnósticas en un intento de tomar decisiones idóneas, contando con una información insuficiente o falseada. No podemos limitar las denominadas pruebas diagnósticas (PD) al espacio de las que se hacen en los laboratorios clínicos, pues ellas, realmente, abarcan aspectos muy amplios, tales como los relacionados con la observación que hace el médico de la ausencia o presencia de signos y síntomas de una

[1] Especialista en Medicina Interna, MgSc en Epidemiología Clínica. Director de la Corporación Académica de Patologías Tropicales, Universidad de Antioquia. Dirección electrónica: cpt_udea@yahoo.com

[2] Profesor de Ginecología y Obstetricia, MgSc en Epidemiología Clínica. Universidad de Antioquia, Medellín, Colombia. Dirección electrónica: jjzuleta@epm.net.co

enfermedad, la práctica de maniobras semiológicas, una tinción especial de un corte histológico, la lectura de una radiografía o de una resonancia nuclear, cambios en potenciales eléctricos, la estimación por ultrasonido de una fracción de eyección, la aplicación de un cuestionario o de unos criterios diagnósticos, etc. La esencia de la práctica de la medicina diagnóstica es el aprendizaje de cuándo se debe aplicar cada elemento de la historia, del examen físico y de la tecnología diagnóstica, además de aprender a ejercer la profesión con la incertidumbre residual que, a pesar del esfuerzo médico, no se logra eliminar del todo. Después de conocer el informe de una PD, el clínico casi nunca sabrá con absoluta certeza si, como en la frase de Epicteto, se trataba de un resultado verdadero o falso, positivo o negativo.

GLOSARIO

Prevalencia: es la proporción de la población afectada por una determinada enfermedad o condición. La prevalencia afecta los valores de predicción (positivos y negativos) de una PD, pero no su sensibilidad ni su especificidad.

Sensibilidad (S): es la capacidad que tiene una PD de identificar correctamente a quien tiene una enfermedad o condición. Es una calidad que sólo se puede determinar en los sujetos enfermos o que tienen la condición que se intenta diagnosticar y por ello se puede afirmar que es el porcentaje de los resultados que son "verdaderos positivos". Su complemento, para tener el 100% del universo de los enfermos, son los "falsos negativos", o sea los sujetos en quienes la enfermedad o condición lamentablemente no se diagnostica, a pesar de tenerla, porque la prueba diagnóstica es negativa de manera equivocada.

Especificidad (E): es la capacidad que posee una PD de identificar correctamente a quienes no tienen la enfermedad o condición que se busca, o sea la calidad de establecer quiénes son "sanos". También se denomina porcentaje de "verdaderos negativos" o "verdaderos sanos". Su complemento, para tener el 100% de los sujetos sanos o que no tienen la condición son los "falsos positivos", o sea los sujetos que no tienen la enfermedad o condición buscada, pero la prueba los identifica erróneamente como afectados.

El valor positivo de predicción (VPP) es la probabilidad de tener realmente una condición o enfermedad, una vez que el resultado de la PD mostró un resultado "positivo". Expresa una probabilidad "condicionada" de gran ayuda para el clínico, porque la probabilidad de padecer la enfermedad está condicionada a la circunstancia de que se practicó la PD y resultó positiva.

El valor negativo de predicción (VPN) es la probabilidad de estar realmente sano, una vez que la PD resultó negativa para la enfermedad.

La razón de probabilidades (RP) o likelihood ratio (LHR) de una PD positiva es un indicador que combina en un solo número los valores de la sensibilidad y la especificidad. La RP positiva de una PD informa la proporción de personas con prueba positiva que están realmente enfermas en relación con las que teniendo la prueba también positiva, en realidad son sanas. La RP positiva indica la relación entre los "verdaderos" positivos (numerador) y los "falsos" positivos (denominador). Ejemplo: una RP de 8 de una PD indica que la prueba tiene ocho veces más probabilidad de resultar positiva entre enfermos que entre sanos.

¿Cómo se aplica la razón de probabilidades (RP)en la clínica? Si la RP es mayor de 10 se puede afirmar, con una gran probabilidad de acertar, que la prueba diagnóstica (signo, síntoma o prueba de laboratorio), cuando está positiva, confirma la enfermedad o el daño. Si la RP es menor de 0,1 la PD niega la existencia de la

enfermedad o el daño. Si la RP está alrededor de la unidad se puede decir que con base en ese estudio no se puede afirmar que la PD sirva de manera concluyente para confirmar o negar la enfermedad que se pretende diagnosticar.

Concordancia entre observadores: es el grado de coincidencia entre quienes leen o interpretan una prueba diagnóstica.

Curva ROC: es una gráfica en la cual se muestran los distintos valores de sensibilidad y especificidad de los puntos de corte de una prueba diagnóstica de orden cuantitativo (colesterol sanguíneo, glicemia en ayunas, hemoglobina, etc.). Las curvas ROC son útiles para observar de manera gráfica cual es el área bajo la curva mayor en los puntos de intersección de la sensibilidad y la especificidad. Así, por ejemplo, para el diagnóstico de diabetes mellitus los expertos podrían escoger distintos puntos de corte: con 90 mg/dl, la sensibilidad para hacer el diagnóstico de diabetes mellitus sería 99% y la especificidad 44%; si se escogiera el punto de corte de 110 mg/dl, la sensibilidad sería 98% y la especificidad 78%; si se optara por un valor de 150 mg/dl, la sensibilidad bajaría al 49%, pero la especificidad subiría al 99%; finalmente, como el punto de corte escogido para el diagnóstico de diabetes fue 125 mg/dl, con una sensibilidad de 78% y una especificidad del 98%, se pueden graficar ambos valores en forma de una curva ROC, siendo el punto de intersección el punto de máximo rendimiento de ambos valores.

Resumen de las características de una prueba diagnóstica: con una tabla de 2 x 2 donde se consigne toda la información alusiva a una prueba diagnóstica se pueden hacer todos los cálculos que sean necesarios para saber cuál es su rendimiento, asumiendo que la prevalencia real de la enfermedad o condición que se estudia está adecuadamente representada en la muestra que se tomó (tabla No.1).

Tabla N° 1
TABLA DE 2X2 CON INFORMACIÓN ALUSIVA A UNA PRUEBA DIAGNÓSTICA

Resultado de la PD	PERSONAS CON LA ENFERMEDAD	PERSONAS SIN LA ENFERMEDAD	TOTALES
POSITIVA	a (Verdaderos positivos) SENSIBILIDAD	b (Falsos positivos)	a + b
NEGATIVA	c (Falsos negativos)	d (Verdaderos negativos) ESPECIFICIDAD	c + d
TOTALES	a + c	b + d	a + b+c+ d

Sensibilidad: proporción de verdaderos positivos = a/(a+c). Cuando una PD de alta sensibilidad está negativa, descarta o niega con gran posibilidad una enfermedad o condición. Las PD de alta sensibilidad se usan frecuentemente en la pesquisa o tamización de una enfermedad, porque la PD tenderá a ser positiva en la mayoría de las personas con la enfermedad o la condición que se busca y unos cuantos casos serán producto de resultados "falsos positivos".

Especificidad: proporción de verdaderos negativos (sanos) = d/(d+c). Una prueba de alta especificidad tiene mucho valor clínico porque cuando es positiva la persona tiene una alta probabilidad de tener realmente la enfermedad o condición que se busca.

Frecuencia de falsos positivos: b/(b+d)=1-especificidad.

Frecuencia de falsos negativos: c/(a+c)=1-sensibilidad.

Prevalencia: proporción de la población afectada por la enfermedad = (a+c)/(a+b+c+d).

Valor positivo de predicción (VPP): proporción de enfermos con la prueba positiva dividida por el total de personas con la prueba positiva. Si la tabla refleja la prevalencia, entonces el VPP = a/(a+b).

Valor negativo de predicción (VNP): proporción de personas sin la enfermedad que se busca que tienen la PD negativa dividida por el total de personas con la prueba negativa. Si la tabla refleja la prevalencia, el VNP = d/(c+d).

Precisión diagnóstica (eficiencia): proporción de resultados correctos = (a+d)/(a+b+c+d).

Razón de probabilidades de una prueba positiva (RP+) = Sensibilidad/(1- especificidad).

Razón de probabilidades de una prueba negativa (RP-) = (1- sensibilidad)/Especificidad.

El profesional que aspire a interpretar de manera adecuada una publicación acerca de PD debe hacerse las siguientes preguntas básicas:

¿Son válidos los resultados del estudio? Los estudios de precisión diagnóstica se ven afectados durante su ejecución por un gran número de variaciones (V), entre las cuales se pueden mencionar: 1) las atribuibles a las características demográficas (una PD puede tener un desempeño diferente en muestras de pacientes que tienen variables demográficas distintas); 2) las dadas por la prevalencia de la enfermedad (la frecuencia de una condición afecta el desempeño diagnóstico de una prueba); 3) las relacionadas con la gravedad de la enfermedad, que pueden afectar la estimación del rendimiento de una PD; 4) las que se dan entre los observadores de las pruebas diagnósticas y también las que son atribuibles a los instrumentos o equipos con los cuales se hace el diagnóstico.

La validez de una investigación acerca de la utilidad de una PD también puede verse afectada por sesgos entre los cuales se encuentran: 1) los de la progresión de la enfermedad (ocurren cuando la PD alternativa se realiza mucho antes de la estándar y esta última se efectúa cuando la enfermedad está muy avanzada); 2) sesgo de tratamiento paradójico que se produce cuando la PD alternativa es de realización anterior al inicio del tratamiento, con aplicación de la prueba estándar cuando ya el tratamiento ha comenzado); 3) sesgo de revisión (la interpretación de una prueba es variada por el conocimiento del resultado de la otra); 4) sesgo de revisión clínica (el conocimiento de datos clínicos afecta la lectura de las pruebas); 5) sesgo de incorporación (el diagnóstico definitivo lo da la prueba alternativa y no la estándar).

Las siguientes preguntas son claves para determinar la validez de una investigación en una prueba diagnóstica:

1. **¿Hubo una comparación "independiente y ciega" con un estándar de referencia?** Es muy importante que a pesar de que la prueba diagnóstica alternativa tenga resultados negativos, se haya aplicado la prueba estándar, especialmente cuando esta última es invasiva o costosa (por ejemplo, una resonancia nuclear magnética), pues la tendencia natural es a no aplicarla si la prueba alternativa fue negativa. Lo de "ciega" quiere decir que quienes están aplicando la prueba alternativa no deben conocer los resultados de la prueba estándar, porque si esta última es positiva, tendrán tendencia a sobreinterpretar la prueba alternativa o a subinterpretarla si la estándar es negativa. Se recomienda no aceptar de manera acrítica la prueba estándar, aun las de tipo histopatológico, pues se sabe, por ejemplo, que las lecturas de biopsias de seno, hígado o piel pueden tener concordancias entre diversos examinadores inferiores al 50%.

2. **¿Incluyó la muestra una variedad apropiada de pacientes, similar a la gama de aquéllos en los que se aplicará la prueba?** La prueba diagnóstica alternativa se debe haber aplicado en un espectro de pacientes similar al que se encuentra en la práctica clínica, o sea, desde pacientes con la enfermedad avanzada y muy evidente, hasta pacientes asintomáticos o con

períodos iniciales de la enfermedad y, además, en pacientes con condiciones confusas o que sean diagnósticos diferenciales de la enfermedad que se está analizando.

Una pregunta no tan importante desde el punto de vista de la validez, pero sí en cuanto a los aspectos operativos y prácticos de reproducir la PD que se analiza, es: **¿Se describen los métodos de desarrollo y aplicación de la nueva prueba con suficiente detalle como para poder replicarla?** Otro aspecto para preguntarse cuando se lee un artículo de pruebas diagnósticas es el siguiente: **¿Cuál es la magnitud del rendimiento diagnóstico de la prueba y cuál es la importancia clínica de los resultados?** La mejor manera de contestarla es mirando si en la publicación se suministran las razones de probabilidad (RP), o se provee la información que se necesita para que el lector la calcule. Finalmente, todas las anteriores preguntas serían inútiles si no recordáramos la preocupación de los autores de la frase con la cual comenzó este escrito: **¿Me ayudará esta prueba diagnóstica que se está analizando a definir mejor la condición de mi paciente y así poder iniciar un tratamiento?**

EJEMPLOS DE ARTÍCULOS DE PRUEBAS DIAGNÓSTICAS

EJEMPLO Nº 1: Review: Medical history, physical examination, and routine tests are useful for diagnosing heart failure in dyspnea.

NOMBRE DE LA PUBLICACIÓN PRIMARIA: Does this dyspneic patient in the emergency department have congestive heart failure? JAMA. 2005; 294:1944-1956.

AUTORES: Wang CS, FitzGerald JM, Schulzer M, Mak E, Ayas NT.

PREGUNTA CLÍNICA: ¿En los pacientes que acuden con disnea a un servicio de urgencias, qué tan útiles son la historia clínica, el examen físico y las pruebas rápidas de laboratorio en el diagnóstico de la falla cardíaca?

REFERENCIA: ACP Journal Club, Marzo/Abril 2006; 144 (2): p 49.

DISEÑO INVESTIGATIVO: revisión sistemática de la literatura de artículos publicados acerca del diagnóstico de falla cardíaca en MEDLINE (de 1966 a julio de 2005) y listados de referencias de artículos relevantes y libros. Se encontraron 22 estudios que llenaban los criterios de búsqueda, pero sólo 18 eran de alta calidad.

CONCLUSIONES: en pacientes adultos con disnea que acuden a un servicio de urgencias, los hallazgos más importantes para diagnosticar una falla cardíaca son, en orden decreciente de valor, los radiológicos (signos de congestión venosa pulmonar y de edema intersticial, ambos con una RP positiva de 12), un hallazgo de un tercer sonido en la auscultación cardíaca (RP positiva de 11), antecedente personal de falla cardíaca (RP positiva de 5,8) y un hallazgo de ingurgitación venosa yugular (RP positiva de 5,2). Los hallazgos que permiten descartar el diagnóstico y buscar otras causas de disnea son, en orden decreciente de valor para negar el diagnóstico, un péptido natriurético tipo B menor de 100 pg/ml (RP negativa de 0,11), la ausencia de cardiomegalia (RP negativa de 0,33), ausencia de congestión venosa pulmonar en la radiografía de tórax (RP negativa de 0,48), ausencia de crépitos en la auscultación pulmonar (RP negativa de 0,51), antecedente negativo de disnea de esfuerzos (RP negativa de 0,48) y antecedentes personales negativos de falla cardiaca (RP negativa de 0,45).

CONFLICTOS DE INTERÉS: no se declaran, por no haber recibido financiación externa.

COMENTARIO DEL EXPERTO: el Doctor Peter C. Wyer del Colegio Americano de Médicos y Cirujanos de la Universidad de Columbia en Nueva York recomienda que para aumentar la utilidad de esta revisión sistemática de las claves diagnósticas de la falla cardíaca en los pacientes con disnea, los médicos deberíamos cotejar cada RP asociada con la presencia o ausencia de un síntoma, signo o hallazgo de laboratorio, con nuestra experiencia personal. Comenta que los hallazgos de esta investigación confirman algunas de las cosas que se saben acerca del diagnóstico de la falla cardíaca: no son tan importantes los antecedentes ni los hallazgos del electrocardiograma, como sí lo son los signos radiológicos en la placa de tórax. Un tercer ruido cardíaco, con una buena razón de probabilidades, a veces es difícil de oír en medio de unas urgencias ruidosas y recomienda que los médicos busquemos más bien la ingurgitación yugular como manera de confirmar la presencia de la falla. Le llama la atención lo poco importante que aparece la disnea de esfuerzos para confirmar el diagnóstico, aunque sí se demostró su valor diagnóstico para negar la falla cuando el paciente no la relataba. Ver ampliación del análisis de este artículo en www.hipertensionyriesgo.com, página de insuficiencia cardíaca congestiva.

BIBLIOGRAFÍA

1. Oxman AD, Sackett DL, Guyatt GH. For the Evidence Based Medicine Working Group. Users guide to the Medical Literature. JAMA 1993; 270: 2093-2095.

2. Lang T, Secic M. How to report statistics in Medicine. BMJ Publishing Group, Philadelphia: American College of Phisicians. 1997.

3. Whiting P, Rutjes A, Reitsma J, Glas A, Bossuyt P, Kleijnen J. Sources of variations and bias in studies of diagnostic accuracy. Ann Intern Med 2004; 140: 189-202.

4. Sackett D, Straus S. Diagnóstico y cribado en: Medicina Basada en la Evidencia, 2ª ed. España: Harcourt Internacional, División Iberoamericana; 2001.

EJEMPLO Nº 2

NOMBRE DE LA PUBLICACIÓN PRIMARIA: Comparación angiográfica de los criterios e índices de alto riesgo para ergometría convencional en pacientes diagnosticados de angina inestable en función del sexo, la edad o el uso de fármacos bradicardizantes.

AUTORES: Alvarez J, Martin E, Romero E, Albadelejo V, De La Hera J, Martin M, Aguado M, Barriales V, Norris C.

AUTOR DE LA LECTURA CRÍTICA: Doctor Wilmar Arley Maya Salazar, Residente de Medicina Interna, Universidad de Antioquia.

PREGUNTA CLÍNICA: ¿En pacientes con angina inestable primaria, la edad, el sexo o la toma de medicamentos que producen bradicardia afectan el rendimiento diagnóstico de la ergometría convencional pronóstica o de las ecuaciones para predecir alto riesgo coronario o del índice de Duke, cuando se comparan sus hallazgos con los de la coronariografía?

REFERENCIA: Rev Esp Cardiol 2006; 59 (5): 448-457.

DISEÑO INVESTIGATIVO: estudio de tipo diagnóstico en el cual se recolectaron 469 pacientes menores de 75 años con diagnóstico de angina inestable, que consultaron de manera consecutiva entre el 1 de enero de 1991 y el 31 de diciembre de 1998. Las características de base de la población informaban de un riesgo cardiovascular importante. Los desenlaces primarios fueron la sensibilidad, especificidad, valores predictivos, razón de probabilidad (RP) de la prueba de es-

fuerzo convencional practicada luego de 48 horas a mujeres, consumidores de fármacos bradicardizantes y pacientes mayores, anginosos, comparados con los hallazgos de la coronariografía.

RESULTADOS: los criterios del Colegio Americano de Cardiología (ACC) y de la Asociación Americana del Corazón (AHA) mostraron tener la mejor sensibilidad (en mujeres:100%, en ancianos: 98,5%, en pacientes bradicárdicos: 96,6%), el mejor valor negativo de predicción (en mujeres: 100%, en ancianos: 96,9%, en pacientes bradicárdicos: 94.3%) y la mejor razón negativa de probabilidades en todos los grupos de pacientes, cuando se quiso predecir el alto riesgo, luego de haber encontrado un descenso importante del ST en la prueba de esfuerzo. Por otro lado, el índice de Duke fue el que mejor especificidad reportó (en mujeres: 90%, en ancianos: 82,1%, en pacientes con bradicardia: 83,8%) y los mejores valores positivos para predicción. El resto de los índices y escalas vieron afectado su rendimiento diagnóstico cuando se trataba de mujeres, ancianos o pacientes con bradicardia.

ANÁLISIS: casi todos los índices para detectar pacientes de alto riesgo en la prueba de esfuerzo han sido establecidos para hombres norteamericanos menores de 65 años; esta investigación representa un esfuerzo para validar la aplicación de los criterios en mujeres, ancianos y pacientes que toman medicamentos que disminuyen la frecuencia cardíaca y que viven en otras regiones del mundo. Usualmente las pruebas cardíacas de esfuerzo tienen una gran variación en su rendimiento diagnóstico, con metaanálisis previos que mostraron una sensibilidad promedio del 75% y una especificidad del 66% para el diagnóstico de enfermedad coronaria grave, en parte como consecuencia de las dificultades existentes en la escogencia de los pacientes. Los autores de la investigación informan que el estudio fue diseñado en los tiempos previos al uso de las troponinas como marcadores de riesgo y eso puede disminuir su utilidad actual. El grupo de investigadores usó la coronariografía como patrón de referencia, prueba que tiene una variación importante entre observadores; en la publicación los autores no aclaran si fue leída por uno o varios hemodinamistas, ni informan si en el momento de la lectura ellos estaban ciegos a los resultados de la prueba de esfuerzo. En cuanto al rendimiento diagnóstico, llaman la atención las bajas RP positivas, la más alta fue para el índice de Duke en mujeres (3,33) y en pacientes con bradicardia (2,34), rendimientos que no fueron significativos. Por último, los investigadores son claros en informar que ellos desean saber si lo reportado en otras latitudes es aplicable a la población española, hecho que nos debe alertar antes de extrapolar los resultados encontrados en poblaciones latinoamericanas. Como conclusión de este artículo se puede señalar que los criterios de la ACC/AHA tienen buen rendimiento diagnóstico cuando se trata de negar la existencia de una enfermedad coronaria grave (si los criterios de las asociaciones están ausentes) y en el caso del índice de Duke, para confirmarla (si el índice es positivo), añadiendo información importante a la interpretación aislada de un segmento ST descendido en una prueba de esfuerzo, en los grupos de pacientes analizados.

EJEMPLO N° 3

TÍTULO: Diagnostic performance of digital versus film mammography for breast cancer screening. (Desempeño diagnóstico de la mamografía digital frente a la mamografía de placas para la tamización del cáncer de mama).

AUTORES: Pisano E, Gatsonis C, Hendrick E, Yaffe M, Baum J, Acharvya S y colaboradores.

REFERENCIA: New Engl J Med 2005; 353 (17): 1773-1783.

RESUMEN: en un programa de tamización de cáncer de mama efectuado en 33 centros de

Estados Unidos y Canadá se realizaron mamografía digital y con placa en un orden asignado al azar, a 49.528 mujeres asintomáticas. Dos radiólogos evaluaron en forma independiente los dos estudios. Emplearon dos escalas de clasificación de las mamografías, la ya validada de BiRads y una segunda diseñada para esta investigación, con siete categorías. El diagnóstico de cáncer se hizo mediante una biopsia dentro de los 15 meses siguientes al ingreso al estudio. A las mujeres cuyas mamografías revelaron lesiones sospechosas se les hizo biopsia y las que no tenían hallazgos anormales se sometieron a seguimiento mamográfico. Se contó para los análisis finales con el 86,3% (42.760) de las mujeres que ingresaron al programa. Los resultados de las prueba se compararon con curvas ROC.

RESULTADOS: la tabla 2 presenta un resumen editado de la tabla 3 del artículo original. Usando la clasificación de BIRADS, 1.249 mujeres (2,9%) tuvieron tanto mamografía digital como de placa positivas, 2.399 (5,6%) tuvieron únicamente mamografía digital positiva, 2.416 (5,7%) tuvieron únicamente mamografía de placa positiva y 36.696 (85,8%) tuvieron ambos estudios negativos. Hubo 335 casos de cáncer confirmados durante los 455 días del estudio extendido y 254 durante el primer año. En la población total, la eficacia de la mamografía digital fue similar a la de la mamografía de placas (la diferencia entre las áreas bajo la curva ROC no fue estadísticamente significativa: 0,03 IC95% - 01,2 a + 0,08). Para los subgrupos planeados con antelación en el protocolo de estudio, mujeres menores de 50 años, mujeres con mamas densas y para las premenopáusicas y perimenopáusicas sí hubo más detección con la mamografía digital (las diferencias entre las respectivas curvas ROC fueron estadísticamente significativas). Con respecto a la sensibilidad y la especificidad, en la siguiente tabla se resumen los resultados a los 365 días de tomada la placa, teniendo en cuenta la clasificación de BIRADS.

Tabla N° 2

	Mamografía digital % (DE)*	Mamografía de placa % (DE)*	Diferencia (IC 95%)
Población total			
Sensibilidad	70 (3)	66 (3)	4 (- 4 a + 12)
Especificidad	92 (1)	91 (1)	0,1 (- 0,3 a + 0,4
Valor predictivo positivo	5 (0,4)	5 (0,3)	
Mujeres menores de 50 años			
Sensibilidad	78 (5)	51 (7)	27 (11 a 44)
Especificidad	90 (3)	90 (3)	0 (- 0,6 a + 0,6)
Valor predictivo positivo	3 (0,5)	2 (0,4)	
Mujeres premenopáusicas y perimenopáusicas			
Sensibilidad	72 (5)	51 (6)	21 (6 a 36)
Especificidad	90 (2)	90 (2)	0,2 (- 0,3 a + 0,8)
Valor predictivo positivo	4 (0,5)	3 (0,4)	
Mujeres con mamas heterogéneas o extremadamente densas			
Sensibilidad	70 (4)	66 (3)	14 (3 a 26)
Especificidad	91 (2)	90 (2)	0,4 (- 0.1 a + 1)
Valor predictivo positivo	4 (0,5)	3 (0,4)	

* Desviación Estándar.

CONCLUSIONES: la exactitud global de la mamografía digital es similar a la de placas, pero es más eficaz para mujeres menores de 50 años, mujeres con mamas radiológicamente densas y para las premenopáusicas y perimenopáusicas.

COMENTARIO: al evaluar la metodología reportada en el artículo, se puede asegurar que se trata de una investigación válida porque el estándar de referencia para confirmar la presencia del cáncer (biopsia) es el más adecuado, se empleó una gama de pacientes en diferentes estadios de la enfermedad en estudio y similar a la población en la cual se emplea la prueba (asintomáticas en una campaña de tamización) y hubo dos evaluadores que interpretaron las mamografías en forma independiente y ciega a los resultados de la biopsia. Aunque no se hizo biopsia a todas, y se pudiera presentar por lo tanto un sesgo de verificación al confirmar únicamente los hallazgos positivos de la

mamografía, el seguimiento clínico para descartar la enfermedad es igualmente aceptable, porque no sería ético exigir biopsia para mujeres con hallazgos negativos.

Con respecto a la interpretación de los resultados, es necesario tener en cuenta varios aspectos. Los investigadores sustentan sus conclusiones a partir de la comparación de las áreas bajo la curva ROC. La curva ROC es una herramienta que grafica la sensibilidad de la prueba en el eje Y, y la resta de 1,0 menos la especificidad en el eje X de un plano cartesiano. Esta curva es útil cuando el resultado de la prueba en estudio no es dicotómico (positiva-negativa) sino que tiene más de dos categorías (específicamente en el BIRADS eran 6 cuando se realizó esta investigación) y por lo tanto requiere un punto de corte en alguna de las categorías para calificar a los individuos como enfermos o sanos. En la medida en que la curva se aproxime más al extremo superior izquierdo del plano, menor es el número de falsos positivos y falsos negativos de la prueba. El área bajo esta curva es una medida objetiva de la eficacia de la prueba y la comparación entre áreas puede diferenciar cuándo una es mejor que otra. Un área de 0,5 bajo la curva indica que la prueba no es adecuada y que es comparable a asignar el diagnóstico de manera aleatoria, y un área de 1,0 sería la de una prueba perfecta. Al comparar el área bajo la curva para dos pruebas se puede identificar cuál de las dos es mejor, y se puede medir la probabilidad de que la diferencia entre ellas se deba al azar; esta probabilidad puede ser alta o baja —valor de p mayor o menor de 5%, respectivamente—. Arbitrariamente se estableció que si esta probabilidad es mayor de 5%, no se puede aceptar que las pruebas tengan eficacias diferentes; sin embargo, este es un criterio netamente estadístico y es posible, como en todo análisis de inferencia, que las diferencias estadísticamente significativas no lo sean clínicamente. Areas de 0,78 y 0,74 bajo las curvas se pueden interpretar como la probabilidad que tiene una mujer que se realice una mamografía, de obtener un resultado acertado con la mamografía digital y de placa, respectivamente. Con respecto a los resultados del estudio, hasta el momento se puede concluir que la eficacia global de ambas pruebas es similar y que en algunos grupos específicos de mujeres hubo diferencias que no las explica el azar.

Al revisar conceptos más conocidos como sensibilidad y especificidad, en la tabla seleccionada como ejemplo se puede apreciar que para la población total existe una alta probabilidad de que las diferencias entre los dos métodos las explique el azar, evidenciado porque el intervalo de confianza de la diferencia entre la sensibilidad y la especificidad de los dos métodos pasa por el cero, que es la igualdad cuando se está evaluando una diferencia de proporciones —diferente a cuando se evalúa el riesgo relativo o el OR en cuyo caso la igualdad es el uno—. Las diferencias de sensibilidad, para los subgrupos analizados, pero no las especificidades, sí son estadísticamente significativas, con unos valores clínicamente importantes —ejemplo 78% frente a 51% para las mujeres menores de 50 años—. Para que estos hallazgos cumplan con una ley básica de las matemáticas, deben llevar inmediatamente a un cuestionamiento: si el método es igual para la población global, pero es superior para unos subgrupos, necesariamente tiene que ser inferior para otros subgrupos; sin embargo, los autores no muestran tal información. Muy posiblemente los grupos en los cuales la mamografía digital sea inferior a la mamografía de placa sean los que tienen las características contrarias a las de los subgrupos en que fue más efectiva.

Otro concepto para evaluar son los valores predictivos positivos: proporción de mujeres que teniendo una mamografía con resultado positivo para cáncer realmente tienen cáncer. En este estudio fluctuó entre 3 y 5%. Esto es un reflejo de la población en que se realizaron los exámenes —mujeres asintomáticas— con muy baja probabilidad de tener cáncer. Los valores

predictivos de las pruebas dependen de la frecuencia de la enfermedad en la población estudiada: para una misma sensibilidad y una misma especificidad de la prueba habrá mayor cantidad de falsos positivos, y por lo tanto, menor valor predictivo positivo si la prueba se aplica en una población con baja prevalencia que si se aplica en una con alta prevalencia de la enfermedad estudiada.

Para casos en los cuales una prueba presenta diferentes niveles de discriminación y con el fin de compensar el efecto que la prevalencia tiene sobre los valores predictivos, actualmente se recomienda que los resultados de las pruebas diagnósticas se presenten en términos de Cocientes o Razones de Probabilidad (likelihood ratio LR), recomendación que no siguieron los autores. Aunque no los presentan, se pueden calcular a partir de los valores de la sensibilidad y la especificidad. En el siguiente artículo se profundizará en este concepto.

En conclusión, al momento de seleccionar una prueba para tamizar, es ideal optar por una con mayor sensibilidad. En este caso, para la población general sería indiferente emplear uno u otro método, pero las mujeres en quienes de antemano se sabe que tienen mamas mamográficamente densas y las premenopáusicas, posiblemente se beneficien del nuevo método.

EJEMPLO Nº 4

TITULO: Second-trimester prediction of severe placental complications in women with combined elevations in alpha-fetoprotein and human chorionic gonadotrophin. (Predicción de las complicaciones placentarias graves del segundo trimestre en mujeres con elevación simultánea de alfa fetoproteína y gonadotropina coriónica).

AUTORES: Alkazaleh F, Chaddha V, Viero S, Malik A, Anastasiades C, Sroka H y colaboradores.

REFERENCIA: Am J Obstetr Gynecol 2006; 194: 821-827.

RESUMEN: el objetivo del estudio fue determinar la capacidad de la evaluación de la arteria uterina con doppler y de la evaluación de la placenta con ultrasonido para identificar resultados clínicos adversos atribuibles a la disfunción placentaria en mujeres con niveles inexplicadamente aumentados de alfa fetoproteína y gonadotropina coriónica humana durante el segundo trimestre. Los resultados adversos estudiados fueron la preeclampsia grave, el bajo peso al nacer, la muerte fetal, el parto antes de las 32 semanas y la ausencia o la reversión del flujo de la arteria uterina al fin de la diástole antes del parto. Se les realizó estudio doppler de la arteria uterina y estudio con ultrasonido de la placenta entre las semanas 19 y 23 de la gestación a 50 mujeres con fetos sin malformaciones y con niveles aumentados de ambos marcadores. Dependiendo de los resultados del doppler y de los hallazgos placentarios, cada mujer recibió un tratamiento diferente (ácido acetilsalicílico, vitamina C, vitamina E, heparina, esteroides) o un esquema de seguimiento paraclínico diferente administrado a discreción de los médicos tratantes. Únicamente el 28% de las placentas fueron normales en el ultrasonido y más del 50% de los doppler fueron anormales. Hubo 13 muertes perinatales (26%), 24 prematuros menores de 32 semanas (48%) y 16 desarrollaron preeclampsia (32%).

RESULTADOS: la sensibilidad del doppler anormal de la arteria uterina sola para detectar anormalidades fluctuó entre 69% (preeclampsia grave) y 100% (muerte intrauterina) con valores predictivos positivos entre 32 y 75%. Las anormalidades simultáneas de la placenta y del doppler tuvieron una sensibilidad del 75% y un LR positivo de 3,3 para predecir el parto antes de 32 semanas, una sensibilidad del 94% y un LR de 3,9 para la restricción del crecimiento intrauterino.

CONCLUSIONES: los autores concluyeron que por ahora este tipo de abordajes diagnósticos se continúen haciendo pero en investigaciones de cohortes de pacientes de alto riesgo.

COMENTARIOS: el presente estudio plantea una utilidad diferente de las pruebas diagnósticas: la predicción. Tradicionalmente la prueba diagnóstica se emplea para confirmar la presencia o ausencia actual de enfermedad. Conceptualmente es difícil aceptar, ante la gran variedad de factores que pueden modificar una condición clínica en un momento dado, que una prueba pueda predecir lo que sucederá con un feto 20 semanas después de realizado el examen. Es necesario asumir las pruebas diagnósticas como herramientas que modifican el nivel de incertidumbre del clínico, antes que ser poseedoras de una verdad absoluta e inmodificable. Una prueba usada con esta intención puede identificar la probabilidad de un grupo de pacientes, pero nunca podrá predecir exactamente lo que sucederá con un individuo en particular, y el objetivo en este caso es alertar al médico para que esté más atento a la evolución de estos individuos de mayor riesgo, sin olvidar que en los grupos de menor riesgo igualmente pueden suceder desenlaces negativos. Un enfoque más razonable para un estudio como éste, es asumir el hallazgo de la prueba diagnóstica como uno más de una serie de factores que pueden modificar el curso de una enfermedad y lo que se quiere evaluar es su influencia en este desenlace.

Asumido en esta forma, el estudio presenta deficiencias importantes en su metodología. Se inició con un grupo heterogéneo de pacientes y hubo cointervenciones terapéuticas; aunque no está probado que algunas de ellas realmente sirvan, pueden llegar a modificar el curso clínico de los pacientes. Es probable que estas diferencias hayan influido en el resultado y por lo tanto no se puede asegurar que el desenlace encontrado se explique exclusivamente por los hallazgos del doppler o el ultrasonido placentario. Ante la situación en que no se cuenta con grupos comparables, la alternativa metodológica recomendada es realizar un ajuste estadístico de todas estas variables de confusión, lo cual no se hizo. El seguimiento de las pacientes fue completo y fue el necesario para poder evaluar el resultado en todas las mujeres incluidas.

Con respecto a los resultados, es importante resaltar que los autores no presentan la precisión de las sensibilidades y especificidades —intervalos de confianza— los cuales probablemente sean muy amplios debido al bajo número de pacientes estudiadas. La otra medida empleada para presentar los resultados es el cociente de probabilidad (LR), cuya mayor utilidad está en las pruebas en que se dan más de dos alternativas de resultado; sin embargo, también se pueden emplear en estos casos. A manera de ejemplo, el LR de 3.1 para muerte fetal intrauterina informa que ante una prueba positiva es 3 veces más probable que la mujer tenga un mortinato a que no lo tenga. Con este LR se puede, a partir de la probabilidad de muerte intrauterina de estas mujeres antes de realizar la prueba, calcular la probabilidad que daría el hecho de tener tal resultado. Los LR modifican en forma importante la probabilidad cuando están en rangos superiores a 10 o inferiores a 0,1; por lo tanto, en este caso la información que aportan (entre 1,4 y 3,9) es poco importante desde el punto de vista clínico.