PUESTA AL DÍA

Métodos de investigación en cardiología clínica (X)

Estudios de evaluación de las pruebas diagnósticas en cardiología

Juan Bautista Cabello López* y Francisco Pozo Rodríguez**

*Unidad de Investigación. Hospital General Universitario de Alicante.
**Servicio de Neumología y Unidad de Investigación. Hospital Universitario 12 de Octubre. Madrid.

El diagnóstico es un proceso cognitivo complejo caracterizado por la incertidumbre, pero gestionable a partir de conocimientos específicos basados en la teoría de la probabilidad. Los estudios de evaluación de tests diagnósticos son el instrumento adecuado para la construcción de ese conocimiento.

Los estudios de diagnóstico pueden considerarse como dos procesos de medida diferentes y un proceso de comparación de ambos. Los procesos de medición (*gold standard* y el test) son independientes y pueden verse influidos por diversas fuentes de variabilidad. Para la comparación entre estándar y test se establece, a priori, una jerarquía explícita en favor del *gold standard*.

Los tests diagnósticos son habitualmente evaluados a través de la sensibilidad, especificidad, cociente de probabilidades y valores predictivos. Sensibilidad (o proporción de verdaderos positivos) y especificidad (proporción de verdaderos negativos) pueden considerarse como proporciones de acuerdo condicionales entre estándar y test, y pueden ser influenciables por el acuerdo debido a azar. En consecuencia los índices *kappa* de sensibilidad y especificidad son instrumentos adecuados para calibrar estos índices.

Sensibilidad y especificidad son estimaciones de los auténticos valores de la población de estudio y por ello pueden presentar errores aleatorios y errores sistemáticos (o sesgos). Los intervalos de confianza de la sensibilidad y especificidad son herramientas matemáticas adecuadas para conocer la precisión de los índices. Un cuidadoso diseño de los estudios es imprescindible para evitar los errores. Finalmente, se expone un catálogo de sesgos a evitar y algunos procedimientos para su corrección.

Palabras clave: *Tests diagnósticos, sensibilidad, especificidad.* Kappa*, intervalos de confianza. Sesgos.*

DIAGNOSTIC TESTS EVALUATION STUDIES IN CARDIOLOGY

Diagnosis is a complex cognitive process wich is characterised by uncertainty. This uncertainty can be managed throught specifical knowledge in conjunction with probability theory. Studies evaluating diagnostic tests are the best way of building this knowledge.

Studies evaluating diagnostic tests have two essential components: the gold standard and the new test. Both components, gold standard and test, are independent measurement process that can be inffluenced by diverse sources of variability. The comparison between diagnostic and test is essentially a hierarchical procedure. Diagnostic tests are evaluated by their sensitivity, specificity compared to a definitive gold standard. The predictive values and the likelihood ratio test are also used.

Sensitivity (the proportion of true positives) and specificity (the proportion of true negatives) are values obtained from a sample and thereby can be considered as the conditional agreement between gold standard and the new test. Kappa coefficients for sensitivity and specificity are ussefull tools for adjusting both indices.

Sensitivity and specificity are non-poblational values, they are estimates of the true values of the study population and can be affected by random error and systematic errors (bias). Conffidence intervals are ussefull for giving an indication of the precision of the point estimates of sensitivity and specificity. A suitable sound design is required to avoiding a biased estimate of sensitivity, likelihood ratio, and predictive values. Finally a list of potential biases is given with methods for minimising these.

Key words: *Test diagnostic studies, sensitivity, specificity. Kappa coefficient, conffidence interval. Bias.*

(*Rev Esp Cardiol* 1997; 50: 507-519)

Correspondencia: Dr. J.B. Cabello López.
Unidad de Investigación. Hospital General Universitario de Alicante.
Maestro Alonso, 109. 03010 Alicante.

«La medicina práctica sería, en suma, la hábil combinación de una verdadera ciencia, la patología que se enseña en los libros, y una gramática parda bondadosa y astuta.»

(Pedro Laín Entralgo)

INTRODUCCIÓN

El diagnóstico es un proceso intelectual complejo que ha sido modelizado de diversos modos[1]. Este proceso exige disponer de un caudal de conocimientos pero su característica fundamental es que se trata de un proceso impregnado de incertidumbre, esa incertidumbre puede ser gestionada a partir de determinados conocimientos basados en la probabilidad[2,3].

En esencia, realizar el diagnóstico es asignar con razonable incertidumbre (es decir, con razonable probabilidad), un paciente a una clase (o grupo) constituida por sujetos con una enfermedad o entidad nosológica. Sobre esa enfermedad disponemos, de una definición más o menos explícita, de conocimientos causales o fisiopatológicos, así como de conocimientos relacionados con el pronóstico previsible o con los tratamientos efectivos. Por lo tanto, el diagnóstico es el primer paso, sin duda crucial, que nos permite la utilización de otras evidencias en la toma de decisiones sobre el paciente.

Esos conocimientos o evidencias sobre pronóstico o sobre tratamiento tienen una indudable orientación práctica, pero se obtienen a partir de modelos formales de estudio desarrollados en otros artículos de esta serie. Por ejemplo, los estudios de cohortes para el pronóstico o los ensayos clínicos para el tratamiento, entre otros.

En el proceso diagnóstico existen determinados pasos que implican una fuerte ganancia de información y son capaces de cambiar (aumentar o disminuir) la probabilidad de pertenencia del paciente a un grupo nosológico[3]. Con frecuencia esas informaciones claves son obtenidas de exploraciones o «pruebas» con algún nivel de sofisticación; sin embargo, algunas preguntas del interrogatorio o determinados signos de la exploración clínica pueden considerarse también como un test diagnóstico en el sentido mencionado[4].

El objetivo del presente artículo es describir los métodos usados para la construcción de conocimiento sobre esos pasos claves denominados «tests» o «pruebas» diagnósticas. Señalemos que al hablar de tests diagnósticos nos referiremos a las informaciones que cumplen el papel señalado anteriormente, es decir, se define test diagnóstico con un criterio funcional.

En otro sentido, el diagnóstico no es un fin en sí mismo, sino un instrumento en la toma de decisiones clínicas; de hecho no es preciso tener una seguridad diagnóstica absoluta para adoptar la decisión terapéutica correcta[5]. El modo más adecuado de utilizar esas pruebas diagnósticas y su combinación con otras informaciones pronósticas o terapéuticas concierne al análisis de decisiones clínicas[6], y será tratado en otro artículo de esta serie.

Al elegir este enfoque hemos tratado de ser coherentes con los objetivos de la serie, orientada a métodos de investigación. Creemos, además, que de este modo respetamos una tradición cardiológica pionera en la incorporación de la teoría de la probabilidad al diagnóstico[7,8].

ESTUDIOS DE EVALUACIÓN DE TESTS DIAGNÓSTICOS

El modelo de razonamiento que subyace en un estudio de evaluación de un test diagnóstico podría esquematizarse del siguiente modo: existe un fenómeno o concepto clínico apoyado por teorías de diversa índole. Se trata, generalmente, de una enfermedad, una manifestación fisiopatológica o clínica de enfermedad, un factor de riesgo, un factor pronóstico. Ese concepto puede ser medido de modo fiable y válido por un procedimiento que se llamará a partir de ahora *diagnóstico* o *gold standard*. Señalemos que al hablar de medición incluimos también la más elemental de todas, la de pertenecer o no a una clase o grupo[9].

En ese marco nos planteamos dos tipos de preguntas; la primera es: ¿hay un segundo procedimiento de medida, que llamaremos test o prueba, que podría medir también ese fenómeno, de modo fiable y válido? La segunda es: ¿ese segundo método tiene algún tipo de ventaja respecto del primero? Las ventajas del segundo método podrán ser teóricas, es decir, que mejore la validez y precisión del primer procedimiento (en cuyo caso estamos buscando un nuevo *gold standard*) o prácticas, es decir, que sea más fácil o económico, con menos riesgo o molestias, etc. (en cuyo caso estamos buscando un procedimiento que evite realizar el *gold standard*).

Ambas preguntas están íntimamente relacionadas: las posibles ventajas prácticas dependen de la validez del test, pero también de otros factores que tienen que ver con la aplicación de valores a las decisiones (valores intelectuales o científicos, económicos, vitales, éticos, etc.). Por su parte, estos factores proporcionan el contexto en el que la primera pregunta adquiere o no pertinencia, es decir, si no obtendremos ventaja alguna previsible, ¿tiene sentido explorar una segunda medida?

En este artículo se asume que el escenario habitual para un estudio de un test diagnóstico es aquel en el que un segundo test puede ofrecer ventajas prácticas sobre el primero, pero nos centraremos en la validez y precisión de la relación entre test y *gold standard*. Se harán, no obstante, algunas alusiones puntuales a la selección de un nuevo *gold standard*.

El esquema conceptual de estos estudios tiene dos niveles, un nivel elemental en el que se realizan dos

procesos de medición distintos e independientes (diagnóstico y test) y un segundo nivel en el que se realiza la validación de una medición respecto de otra que consideramos de superior jerarquía. Puede, de algún modo, considerarse como un caso particular de los procesos de validación (validación por criterio).

La estrategia de investigación en la evaluación de un test diagnóstico consiste en seleccionar de una población una muestra de pacientes, aplicar en ellos el «diagnóstico» y el «test» problema, y estimar unos descriptores básicos de la relación entre ambos y ciertas combinaciones de ellos.

Algunos aspectos de la arquitectura de los estudios plantean diferentes efectos sobre la validez de los descriptores básicos y de sus combinaciones. Por ello, se abordarán en primer lugar los procesos de medición. Posteriormente, analizaremos el proceso de validación de una medida respecto de otra, destacando cuáles son las características de calidad exigibles a los descriptores. Finalmente, se tratará de las diferentes arquitecturas de estudio y su impacto sobre la validez y precisión de las mediciones y descriptores.

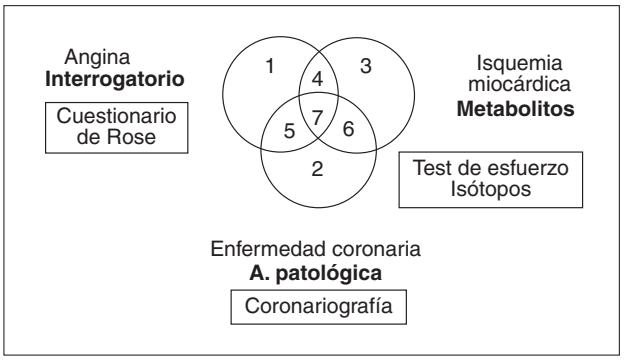

*Fig. 1. Representación en diagramas de la relación conceptual entre angina de pecho, isquemia miocárdica y enfermedad aterosclerosa (tomada y modificada de Hlatky)[11]. 1 y 5: dolor torácico de origen no cardíaco; 3: hipoxemia, estenosis aórtica, etc.; 4: vasospasmo coronario puro; 2: enfermedad coronaria asintomática; 6: isquemia silente; 7: síndrome clásico. En negrita se exponen los tests óptimos para la identificación de cada uno de los conceptos (*gold standard *auténticos), en recuadro los tests usados habitualmente como* gold standard *(*gold standard *prácticos).*

Procesos de medición. El diagnóstico y el test

Cualquier medición tiene dos componentes básicos, un concepto o *constructo* a medir, y un procedimiento del que cabe destacar tres elementos: el sistema de reglas o protocolo para proceder a la medición, el resultado expresable en alguna escala y los criterios de interpretación. La calidad del proceso de medida puede ser evaluada a través de las dos características esenciales que señalamos en otro artículo de esta serie[10] (precisión y validez) que serán exigibles tanto al diagnóstico como al test. Existen, además, otras consideraciones a realizar en relación con el *gold standard* y con el test.

El diagnóstico (gold estandard)

Es el procedimiento que permite medir de modo fiable y válido un determinado concepto clínico, habitualmente la existencia de enfermedad. Las enfermedades son entidades de conocimiento (nosológicas) sustentadas por evidencias empíricas, pero con alto nivel de teorización y frecuentemente con diferentes formas clínicas.

Consideremos, por ejemplo, tres *constructos* teóricos distintos[11]: «enfermedad coronaria aterosclerosa», que es un concepto consagrado por múltiples teorías y observaciones, «angina de pecho» cuya descripción por Heberden se remonta al siglo XVIII, o «isquemia miocárdica» que es un concepto fisiopatológico. Son, respectivamente, un concepto anatómico, un concepto clínico y un concepto funcional, y representan diferentes paradigmas sobre la enfermedad. Es obvio que los tres conceptos están muy correlacionados, aunque no sean exactamente lo mismo. De hecho, sabemos que puede haber angina sin enfermedad coronaria, isquemia sin angina, isquemia sin enfermedad coronaria, etc.(fig. 1).

Cada uno de estos conceptos puede ser medido de manera óptima por diferentes procedimientos, que incluyen un sistema de reglas, un resultado y unos criterios de interpretación. Así, por ejemplo, para la enfermedad coronaria sería la anatomía patológica, con su correspondiente procesado, la observación por un patólogo y la emisión de un juicio sobre la existencia de arteriosclerosis. Para la angina de pecho sería un juicio clínico realizado por un experto cualificado (o un panel de expertos) tras un interrogatorio clínico. Para la isquemia miocárdica el método teóricamente óptimo sería la realización de determinaciones metabólicas en seno coronario, su procesado en laboratorio y su posterior interpretación.

El proceso podrá fraccionarse, por ejemplo, en segmentos coronarios, y podrá ser subsidiario de mayor o menor instrumentación tecnológica[12,13]. Cada una de estas mediciones podrá expresarse en diferentes escalas (binarias, ordinales, cuantitativas, etc.).

A efectos de este artículo se considerará que el resultado de la medición del estándar se expresa como positivo o negativo, que en el contexto clínico habitual correspondería a la pertenencia al grupo de enfermedad o no enfermedad.

En cualquier caso, si se acepta que esas mediciones son óptimas, es porque existe poca variabilidad en los juicios (medida a través de los indicadores *kappa* o equivalentes), y porque existen otros conocimientos que sustentan la validez del proceso. Por ejemplo, para el caso de la enfermedad coronaria existe poca variación entre diferentes patólogos, y disponemos observaciones de arterias normales, conocimiento de

la fisiología cardíaca, correlaciones anatomoclínicas o conocimientos experimentales.

Es decir, podemos comprobar empíricamente su repetibilidad o consistencia, pero aceptamos su validez clínica en función de un marco de conocimiento que poseemos y existe un amplio consenso en que es la mejor aproximación a ese constructo o fenómeno en ese marco tecnológico y conceptual. Así pues, resulta casi superfluo ponderar la importancia de una definición explícita del concepto que tratamos de medir, y señalar que el *gold standard* será transitorio y podrá ser sustituido al mejorar nuestro conocimiento teórico o nuestras posibilidades tecnológicas.

Por razones diversas, pocas veces podemos usar esas mediciones óptimas en la clínica. Así, usar la anatomía patológica coronaria como estándar no es posible por razones obvias; lo mismo cabe decir para las determinaciones metabólicas. En consecuencia necesitamos buscar otros procedimientos más factibles para medir enfermedad coronaria o isquemia miocárdica, que puedan actuar como estándar en la práctica y en la investigación clínica.

Estos otros procedimientos sustitutivos deberán ser también fiables y válidos, es decir, deben mostrar resultados reproducibles y un alto grado de correlación cuando se comparan con los anteriores, según los modelos de comparación de los que se tratará posteriormente. En nuestro caso, por ejemplo, podríamos ver la consistencia en la interpretación de las coronariografías y posteriormente enfrentar la coronariografía a la anatomía patológica valorada tiempo después en pacientes fallecidos[12] (validación predictiva). De ese modo podemos decidir si es posible usar la coronariografía como estándar para la enfermedad coronaria. Análogo planteamiento podría realizarse para usar la escintigrafía con talio como estándar de la isquemia miocárdica en vez de las determinaciones metabólicas. En ambos casos puede aceptarse que, aunque no son las mediciones óptimas del fenómeno, pueden funcionar como *gold standard* en la práctica y en la investigación.

Así pues, se puede considerar que el *gold standard* es un papel a desempeñar por una medida, a la que cabe exigir precisión, validez y otras cualidades relacionadas con la factibilidad y los valores. En ocasiones es imposible disponer de un estándar por la naturaleza del concepto a medir o por la ausencia de conocimiento suficiente, en esos casos un recurso a considerar puede ser el explorar la «validez consensual» a partir de un panel de expertos o de consensos más amplios. El caso de la angina de pecho[11] o los sucesivos criterios de Jones[14,15] para la fiebre reumática pueden ser interesantes ejemplos de este proceder.

Aun así, en ocasiones sólo se dispone de un estándar imperfecto, que puede provocar errores en el resultado del estudio *(sesgo del estándar imperfecto)*. En estos casos puede ser de interés realizar el estudio y corregir ese sesgo a partir de modelos matemáticos[16-18].

El test o prueba en estudio

Es la prueba que podría sustituir con alguna ventaja práctica al estándar. En unas ocasiones el interés es usar un método más fácil, más económico, o evitar riesgos a las personas (p. ej., usar test de esfuerzo en vez de estudio isotópico). Otras veces la imposibilidad de usar un determinado estándar depende del contexto en el que se va a usar, por ejemplo, sería complicado emplear a un panel de expertos para diagnosticar la angina de pecho en un estudio poblacional en el que cientos de personas van a ser incluidos. Para esos fines es más eficiente usar el cuestionario de Rose[19], estudiando previamente su consistencia y su validez[20] (indirectamente o frente a un panel de expertos). En los dos ejemplos citados, el constructo medido por el test es el mismo que el medido por el estándar (isquemia en un caso y angina en otro) y, puesto que se trata de dos modos distintos de medir el mismo concepto, no se plantea ningún conflicto lógico.

Sin embargo, en la investigación clínica se comparan frecuentemente dos medidas de conceptos distintos, aunque íntimamente relacionados. Por ejemplo, podemos usar como *gold standard* la coronariografía y como test la prueba de esfuerzo. O incluso se pueden comparar dos pruebas entre sí a través de un estándar, por ejemplo la ecocardiografía de estrés y el test de esfuerzo usando como estándar la coronariografía[21]. En estos casos medimos cosas diferentes y, por tanto, existe una laguna lógica en el proceso de medición. Aun así, existen ventajas prácticas de realizar esa comparación, derivadas de la dinámica del proceso de diagnóstico. En consecuencia, aunque señalemos ese conflicto lógico, lo que cabe exigir al test y al estándar es que tengan una intensa relación conceptual en sus constructos, e idealmente que midan el mismo constructo.

Los otros elementos del procedimiento de medida son el sistema de reglas, los resultados o respuesta y los criterios de interpretación.

El sistema de reglas de medición *o protocolo* es un conjunto de normas operativas que fijan de manera concreta las condiciones de realización de la medición: pueden diferenciarse dos tipos de protocolos en función del tipo de variable que tratemos de medir en la respuesta[22]. En unos casos se trata de medir una característica, condición o síntoma del paciente *(índices de estado)*, por ejemplo, el electrocardiograma basal, el nivel de creatincinasa o la existencia de una imagen tomográfica. En otros casos, muy frecuentes en cardiología, se trata de observar la respuesta funcional del paciente ante un estímulo *(índices estímulo-*

respuesta). El test de esfuerzo, el test de esfuerzo con talio, o la ecocardiografía con dobutamina o dipiridamol serían ejemplos de estos índices. Se comprende que la complejidad del protocolo será superior en el segundo caso, ya que para los índices de estado deben fijarse las condiciones de realización, mientras que en los índices estímulo respuesta deben fijarse además las condiciones del estímulo.

La *respuesta* o resultado es la información observada en el paciente. Cuando esa información se refiere a un índice de estado se expresa como un valor en una escala determinada. Por ejemplo, escala binaria (positivo/negativo) o escala dimensional numérica (255 U/l de CPK). Cuando se trata de un test estímulo-respuesta el resultado se expresa en términos de cambio (descenso de ST, defecto o cambios de perfusión, etc.). De manera general, las escalas numéricas serán preferibles a las ordinales o a las binarias puesto que contienen más información. Cuando la respuesta se ofrece en escalas numéricas, el análisis se optimiza si se aplican las curvas de características operacionales del receptor[23] (curvas ROC), o los cocientes de probabilidad *(likelihood ratio)*[4]. Por razones de simplificación, en el resto del artículo se considerarán los resultados del test simplemente en escala binaria (positivos o negativos).

Frecuentemente el resultado no es una variable única sino un conjunto de variables que además pueden tener complejas relaciones entre sí. Por ejemplo, en el test de esfuerzo son variables de interés la duración del esfuerzo, la aparición de síntomas, la frecuencia cardíaca, el patrón de aumento de la frecuencia, los cambios en la morfología del segmento ST, los milímetros de descenso del ST, los cambios en la R, el comportamiento de la presión arterial, etc.[24-26]. Por tanto, podemos considerar que la respuesta puede referirse de forma simultánea a diveras dimensiones, y se puede definir matemáticamente la respuesta como un vector *X1* en un espacio real multidimensional.

La *referencia* es un conjunto de criterios por los que se clasifica cada resultado observado como positivo o negativo. Para el caso de un conjunto de variables interrelacionadas (multidimensional), puede ser de interés proceder al análisis del test según subtipos de respuesta, o mejor el uso de métodos de integración a partir de modelos multivariantes[27].

En resumen, tanto diagnóstico como test son procesos de medida que exigen clarificación conceptual, y que están influidos (o amenazados) por múltiples fuentes de variabilidad[28]. La función de protocolo, de las escalas para medir la respuesta, de los criterios de interpretación y del entrenamiento específico es disminuir la variabilidad de la medición, es decir, hacerla precisa o consistente. La validez del *gold standard* es un proceso complejo basado en una red de argumentos y la validez de test se explorará comparándolo con el diagnóstico.

Comparación de mediciones

Modos de comparación

Realizadas las dos mediciones básicas, diagnóstico y test, debe valorarse la relación entre ambos usando dos tipos de comparación diferentes.

Comparación abierta. En primer lugar, se procede a evaluar la relación entre diagnóstico y test sin condiciones previas, es decir, explorar la existencia de «asociación» estadística entre diagnóstico y test, mediante algunas de las técnicas de correlación.

Si se observa correlación entre el test y el diagnóstico, puesto que este último es un indicador válido de enfermedad, podremos concluir que el test también es indicador de enfermedad. Si, por el contrario, no se observa asociación entre diagnóstico y test, es decir, existe independencia estadística, no procede continuar la evaluación.

Por ejemplo, se puede buscar la asociación entre la existencia enfermedad coronaria y elevación sérica de la fosfatasa ácida, o también la correlación entre enfermedad coronaria angiográfica y comportamientos anómalos en el test de esfuerzo. Probablemente no se obtenga éxito en el primer caso y sí se encuentre que la proporción de enfermos con anomalías del test de esfuerzo es mucho mayor en los sujetos con enfermedad coronaria respecto de los que no la tienen. Sin embargo, la existencia de esa asociación entre enfermedad coronaria y anomalías en el test de esfuerzo no es útil para saber si la enfermedad coronaria puede diagnosticarse por un test de esfuerzo anormal. Si se desea usar el test de esfuerzo para el diagnóstico de enfermedad coronaria es necesario que exista asociación pero no es suficiente, hay que profundizar en esa relación comparando los procedimientos de otro modo.

Comparación jerárquica. En segundo lugar, se comparará el diagnóstico y el test, asumiendo que el diagnóstico es el indicador válido de la existencia o no de enfermedad. Esta segunda es una comparación en la que existe una jerarquía explícita en favor del diagnóstico. Este aspecto tiene interés por cuanto implica que, con este modelo de comparación que planteamos, es lógicamente imposible que un test, por nuevo y prometedor que sea, supere al *gold standard*. Por tanto para seleccionar un nuevo *gold standard* deben considerarse otros métodos de validación de la medida.

Ambas comparaciones, abierta y jerárquica, requieren que se cumpla la asunción de que las mediciones son independientes en dos sentidos: por una parte, los métodos de medir el diagnóstico y el test no pueden tener elementos comunes (en jerga escolástica «lo definido no puede entrar en la definición»). Por ejemplo, no es razonable explorar la asociación de la corea mi-

TABLA 1
Interpretación de los valores de *kappa*

Valor de *kappa*	Grado de concordancia
0,81-1,00	Excelente
0,61-0,80	Buena
0,41-0,60	Moderada
0,21-0,40	Ligera
< 0,20	Mala

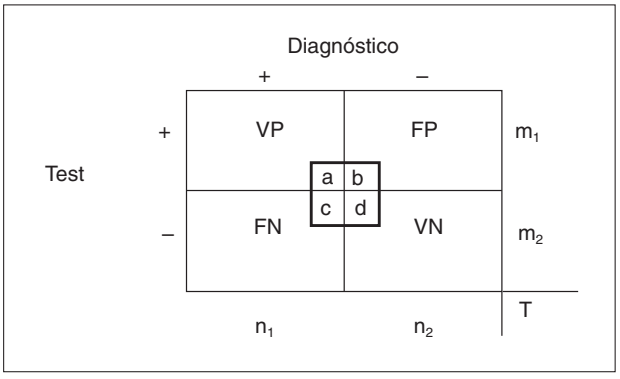

Fig. 2. Comparación de diagnóstico y test. Enfermos: n1, no enfermos: n2, test positivo: ml, test negativo: m2, Total: T. VP y a: verdaderos positivos, FN y c: falsos negativos, FP y b: falsos positivos, VN y d: verdaderos negativos.

nor y fiebre reumática (que obviamente existirá) y tampoco su valor para el diagnóstico. La inclusión de elementos de la definición de enfermedad en el test es un error conceptual[29] que recibe la denominación de *sesgo de incorporación*.

Por otra parte, los dos procesos de medición (diagnóstico y test) deben realizarse de modo «ciego» para preservar a la medición de la existencia de posiciones a priori por parte del observador, influidas por el conocimiento de la otra medida. Esta precaución es particularmente importante cuando el resultado se basa en juicios de los observadores. Podemos imaginar fácilmente que la lectura de una radiografía de tórax puede ser bien distinta en función de que sepamos o no que el paciente tiene insuficiencia cardíaca, o que la interpretación de imágenes observadas y el esfuerzo diagnóstico en la búsqueda de verrugas ecocardiográficas será superior si sabemos que el paciente tiene endocarditis (o reinterpretamos la ecocardiografía tras recibir los hemocultivos).

Así pues, la ausencia de cegado en la evaluación del diagnóstico o del test reciben, respectivamente, el nombre de *sesgo por revisión del diagnóstico* o *sesgo por revisión del test*.

Descriptores básicos y combinaciones de ellos

Los descriptores básicos[30,31] usados se exponen en la tabla 1, y a continuación se describen. Del total de pacientes (T), unos serán enfermos según el estándar (n1) y otros no enfermos (n2) y, asimismo, del total T, m1 tendrán el test positivo y m2 el test negativo. Los cruces entre los resultados del test y del diagnóstico ofrecen cuatro descriptores que corresponden a las posibles situaciones posibles. Cualquier paciente puede ser clasificado como verdadero positivo (VP, a) si padece la enfermedad de acuerdo con el *gold standard* y el resultado del test es positivo, falso positivo (FP, b) si no padece la enfermedad y el test es positivo, falso negativo (FN, c) si padece la enfermedad el resultado de test es negativo, y verdadero negativo (VN, d) si no padece la enfermedad y el resultado del test es negativo.

Cálculos combinados

Para la comparación jerárquica o para el uso clínico interesa conocer dos grupos de indicadores obtenidos a partir de la figura 2. Todos ellos son proporciones que serán interpretables como probabilidades de que determinados episodios ocurran en función que de otros se hayan producido (probabilidades condicionales).

1. Cálculos verticales. Si aceptamos que el diagnóstico define la existencia de enfermedad pueden calcularse dos índices: sensibilidad (Se), proporción de verdaderos positivos, o probabilidad de que el test sea positivo dado que el diagnóstico es positivo y especificidad (Es), proporción de verdaderos negativos, o probabilidad de que el test sea negativo dado que el diagnóstico es negativo. Los complementarios serían, para la sensibilidad la proporción de falsos negativos (1 – Se) y para la especificidad la proporción de falsos positivos (1 – Es).

2. Cálculos horizontales. Si conocemos el resultado del test pueden calcularse otros dos índices: valor predictivo positivo, proporción de enfermos entre los tests positivos, o probabilidad de enfermedad dado que el test es positivo y valor predictivo negativo, proporción de no enfermos entre los tests negativos, o probabilidad de ser no enfermo dado que el test es negativo.

Los cálculos verticales informan sobre las características del test asumiendo la jerarquía del diagnóstico y se refieren a la validación por criterio del test, son índices guiados por la nosología. Los cálculos horizontales informan sobre las consecuencias probabilísticas de un test positivo o negativo (que es la pregunta que procede en actividad clínica), son índices guiados por el test. Los indicadores de ambas columnas se relacionan entre sí a través de la fórmula de Bayes.

Valor predictivo positivo:

$$VP+ = \frac{Se \times prevalencia}{Se \times prevalencia + (1 - Es) \times (1 - prevalencia)}$$

Valor predictivo negativo:

$$VP- = \frac{Es \times (1 - prevalencia)}{Es \times (1 - prevalencia) + (1 - Se) \times prevalencia}$$

Al observar las características de pruebas diagnósticas resulta obvio cuáles son más sensibles o más específicas, pero no es posible saber qué pruebas tienen una mejor combinación de sensibilidad y especificidad. El problema de la combinación de sensibilidad y especificidad se soluciona a través de otro índice denominado cociente de probabilidades *(likelihood ratio)*.

Se define el cociente de probabilidades positivo como cuánto más probable es que una prueba sea positiva en un paciente con enfermedad respecto de uno que no la tenga.

$$CP+ = \frac{Se}{(1 - Es)}$$

El cociente de probabilidades negativo se define como cuánto más probable es que una prueba sea negativa en un paciente con enfermedad respecto de uno sin enfermedad.

$$CP- = \frac{(1 - Se)}{Es}$$

El cociente de probabilidades es cada vez más usado en la clínica puesto que ofrece una información combinada de los índices verticales y ofrece, además, algunas otras ventajas cuyo análisis supera los objetivos de este artículo.

Otros aspectos de los estimadores combinados

Estimación por intervalos de Se y Es. Sensibilidad, especificidad y valores predictivos son proporciones que hemos obtenido de un muestreo y no de una población. Son, por tanto, una estimación puntual de esas proporciones poblacionales, que por definición nos son desconocidas. Cabe preguntarse cuánto podrían oscilar por azar dichas proporciones si con el mismo criterio de selección hubiéramos obtenido otra muestra de igual tamaño pero diferentes elementos, o de otro modo cuál es el error aleatorio de esas estimaciones. Esta pregunta puede contestarse realizando una estimación por intervalo a partir del cálculo del error estándar de las proporciones[32] usando la distribución binomial, o si n1 y n2 son suficientemente grandes la aproximación normal a la binomial*.

Para la Se el intervalo de confianza será:

$$Se \pm 1{,}96 \sqrt{\frac{Se \times (1 - Se)}{n_1}}$$

siendo $\sqrt{\frac{Se \times (1 - Se)}{n_1}}$ el error estándar de la sensibilidad.

Para Es $\quad Es \pm 1{,}96 \sqrt{\frac{Es \times (1 - Es)}{n_2}}$

siendo $\sqrt{\frac{Es \times (1 - Es)}{n_1}}$ el error estándar de la especificidad.

Coeficientes kappa *de Se y Es.* En realidad al referirnos a sensibilidad y especificidad estamos hablando de concordancia o acuerdo condicional[29]. La sensibilidad es el porcentaje de acuerdo entre test y diagnóstico dado (o condicionado) que el estándar es positivo. La especificidad será el porcentaje de acuerdo entre diagnóstico y test dado que el diagnóstico es negativo.

Pero además sensibilidad y especificidad son medidas «no calibradas» y realmente no oscilan entre 0 y 1 sino que el nivel de comienzo depende de la proporción de test positivos por azar en la población estudiada para la sensibilidad o de la proporción de test negativos por azar para la especificidad[33,34]. Por ejemplo, en el caso de la relación entre enfermedad coronaria y fosfatasa ácida, aunque no exista asociación, encontraremos por azar algunos pacientes en cada una de las cuatro casillas de la tabla 1. Por tanto, cabe preguntarse ¿qué porcentaje del acuerdo condicional (Se o Es) es debido al azar (sería observable incluso si diagnóstico y test fueran independientes) y qué porcentaje se debe a la capacidad del test para clasificar enfermedad?

Si estándar y test no tuvieran asociación alguna (fueran estadísticamente independientes), cabría esperar que la proporción de tests positivos entre los enfermos fuera igual que la que se observa en el total de la muestra T, e igual, a su vez, a la que se observa en los no enfermos. En nuestro ejemplo anterior, la proporción de personas con elevación de fosfatasa ácida será igual entre los enfermos y entre los sanos. Podemos calcular el valor esperado por azar para la casilla a

*Las fórmulas para calcular el intervalo de confianza de sensibilidad y especificidad dependen del tipo de muestreo que se haya realizado (véase posteriormente). La fórmula que exponemos es la correspondiente al diseño retrospectivo.

(asumiendo independencia) a partir de una regla de tres simple, será:

$$E(a) = \frac{(a + b) \times (a + c)}{T}$$

Es decir, el número de efectivos de la casilla «a» (enfermos coronarios con aumento de fosfatasas ácidas) dependerá de la prevalencia de enfermedad coronaria en nuestra muestra (a + c) y de la proporción de personas con elevación de fosfatasa en nuestra muestra (a + b).

En consecuencia el valor esperado de la sensibilidad será el cociente del valor esperado de la casilla «a» y el número enfermos (a + c), luego sensibilidad esperada por azar:

$$Se(E) = \frac{\frac{(a + b) \times (a + c)}{T}}{a + c} = \frac{a + b}{T}$$

Nótese que a + b/T es la proporción de test positivos en el total de la muestra que denominamos P, su complementario (1 – P), o también P', será la proporción de tests negativos en la muestra (c + d/T).

Se define el índice «*kappa* ponderado de la sensibilidad» como el cociente entre la sensibilidad observada no atribuible al azar y la máxima sensibilidad observable no atribuible al azar.

K = Se (observada) – Se (esperada)/1 – Se (esperada)

$$\text{luego } K_{Se} = \frac{Se - \frac{a + b}{T}}{1 - \frac{a + b}{T}}$$

$$\text{Simplificando } K_{Se} = \frac{Se - P}{1 - P}$$

Procediendo de modo análogo con la especificidad. Se obtiene:

$$K_{Es} = \frac{Es - \frac{c + d}{T}}{1 - \frac{c + d}{T}}$$

$$\text{Simplificando } K_{Es} = \frac{Es - P'}{1 - P'}$$

Ambos *kappas* informan de la calidad de los indicadores de Se y Es, que serán próximos a 1 si la Se (o Es) atribuible al azar es pequeña y próximos a 0 si es grande. Estos índices *kappa* son contrastables con la hipótesis nula de K = 0, es decir podemos comprobar si los *kappas* son significativamente distintos de cero. Asimismo podemos evaluar cómo de buenos son los *kappas* obtenidos para los indicadores descritos. A tal efecto pueden usarse los criterios de Landis y Koch[35] (tabla 1).

Arquitectura de los estudios de diagnóstico

En estos estudios el objetivo es conocer unos descriptores que puedan ser utilizados en el proceso de asignar probabilísticamente un paciente a una categoría diagnóstica determinada. El método de investigación que usamos se basa en estimar esos índices a partir de muestras de pacientes suficientemente representativas.

Por ello, existen determinados aspectos en el diseño que pueden afectar a la precisión (errores aleatorios) o la validez (errores sitemáticos o sesgos[36,37]) de las estimaciones realizadas, y de este modo pueden amenazar la calidad de la evidencia que aporta el estudio.

A continuación se exponen los conceptos clave del diseño, se completan los conceptos de precisión introducidos a propósito del intervalo de confianza, y se describen en el texto y tabla 2, los posibles sesgos de estos estudios y los mecanismos para su control (tabla 2).

Población diana de estudio

La definición de la población es uno de los problemas importantes a efectos de generalizar los resultados de la evaluación de un test diagnóstico, y en consecuencia, a la hora de utilizar los resultados en una situación clínica concreta. De manera general, la población de estudio deberá parecerse al escenario clínico en el que funcionará la prueba. Sin embargo, un test puede usarse en personas con alta o baja sospecha de enfermedad, en una subpoblación determinada de pacientes con una enfermedad, o incluso puede tratarse de población sana en la que el objetivo es la detección precoz de enfermedad, o de factores de riesgo. Muchas de esas situaciones difieren fundamentalmente en la prevalencia de la enfermedad cuyo test diagnóstico estamos estudiando, pero pueden existir otros factores a considerar.

Clásicamente se asumía que la sensibilidad y especificidad (cálculos verticales) son independientes de la prevalencia, y que los valores predictivos positivo y negativo reales podrían corregirse para cada prevalencia a través de las fórmulas de Bayes. Sin embargo, la

TABLA 2
Sesgos en estudios de tests diagnósticos

Tipo de sesgo	Modo de producción	Consecuencias	Modos de control
1. Sesgo por inadecuado espectro de enfermedad, o sesgo de selección de casos[36-39,42]	No se tiene en cuenta el espectro clínico, patológico o de comorbilidad	Sobreestima Se y Es si se representa a los casos graves y si se trata de casos leves infraestima Se y Es	*1.* Representar el espectro completo en la muestra *2.* Describir el espectro en el análisis *3.* Análisis del test en los subgrupos
2. Sesgo del *gold standard* imperfecto[16-18]	No se dispone de un buen *gold standard* y se usa el disponible (aunque no clasifique muy bien)	Generalmente sobreestima Se y Es A veces infraestima Se y Es	*1.* Seguimiento clínico de los pacientes para ver si son enfermos o no *2.* Correcciones matemáticas si se dispone de un subconjunto de pacientes con una adscripción definitiva
3. Sesgo de incorporación[29,36,37]	Elementos del test forman parte del *gold standard* (están incorporados)	Sobreestima Se y Es	Conceptualización adecuada del *gold standard* y de test
4. Sesgo por revisión del diagnóstico o del test[29,36,37]	La interpretación del test o del estándar se realiza conociendo el otro resultado, es decir de modo *no ciego*	Sobreestima Se y Es	Cegado de las personas que interpretan (y realizan) el estándar y el test
5. Sesgo de verificación diagnóstica[36,37,44,45]	El resultado del test condiciona la realización del *gold estándar*	Sobreestima Se e infraestima Es	*1.* Realizar estándar en todos los pacientes del estudio. *Si no es posible hacerlo* *2.* Seguimiento de test negativos *3.* Correcciones matemáticas
6. Resultados no interpretables[36,37,43]	Es una eventualidad que se produce en cualquier test o estándar	Sobreestima Se y Es	*1.* Repetición del test, si es posible *2.* Inclusión en el análisis de los casos no interpretables
7. Sesgo por variabilidad en la interpretación de resultados[36,37]	Diversos observadores que actúan dentro del estudio tienen diferentes Se y Es El mismo observador cambia su Se y Es dentro del estudio por el entrenamiento	Generalmente infraestima la Se y Es	*1.* Estudios previos (piloto) de consistencia interobservadores *2.* Correcciones matemáticas

asunción se ha demostrado empíricamente errónea[29,36,37-39], de modo que la sensibilidad y la prevalencia pueden estar relacionadas en función del espectro de enfermedad.

Si se estudia a un grupo de sujetos con enfermedad coronaria avanzada la capacidad del test de esfuerzo para diagnosticar enfermedad (sensibilidad) será mayor que en pacientes con enfermedad coronaria menos avanzada. De este modo, si se selecciona una serie de pacientes hospitalarios, en los que la prevalencia es mayor y la enfermedad más grave, la sensibilidad obtenida del test estará sobreestimada, y hubiera sido menor si los pacientes se hubieran obtenido de la consulta de un centro de salud. En el primer caso se estaría estudiando el funcionamiento del test en la fracción más avanzada del espectro de enfermedad y en el segundo en la menos avanzada.

Por otra parte, diferentes formas clínicas de enfermedad pueden presentar distintas sensibilidades y especificidades para determinados tests[11]. Si revisamos la figura 1 se comprenderá las diferentes Se y Es del test de esfuerzo en las diferentes formas clínicas. También determinadas localizaciones patológicas de la enfermedad pueden procurar distintos valores en las pruebas como ocurre con el test de esfuerzo y las diferentes localizaciones de la obstrucción.

Otras características del paciente y particularmente la comorbilidad pueden también afectar a los índices verticales e incluso la fiabilidad de las mediciones básicas. Ejemplos de ello serían el sexo, la diabetes o la presencia de fármacos para la sensibilidad y especificidad del test de esfuerzo o de la escintigrafía[21,40,41], y también el sexo para la fiabilidad del cuestionario de Rose[20].

Por tanto, en la definición de la población y en la selección de la muestra debe ponerse atención cuidadosa en la representación del espectro completo de la enfermedad, considerado éste desde perspectivas clínicas, patológicas y de comorbilidad[36-39] *(case mix)*. En todo caso, la descripción del *case mix* en los estudios de diagnóstico debe ser completa, y en el análisis deben probarse técnicas multivariantes para conocer el rendimiento del test en los correspondientes subgrupos.

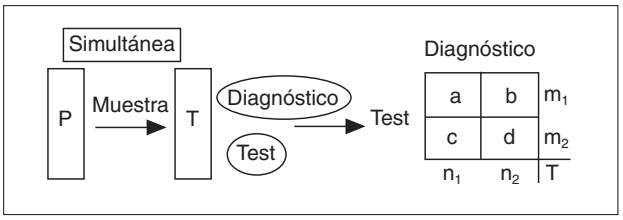

Fig. 3. Estrategias de muestreo en los estudios de evaluación de tests diagnósticos: simultánea; P: población de estudio; T: muestra.

Estrategias de muestreo

«Estrategia o diseño simultáneo.» En la que se procede del siguiente modo (fig. 2): se toma una muestra de tamaño T de la población diana del estudio, a la que se le aplican el diagnóstico (D) y el test (T) de modo concurrente guardando los requisitos de cegado para preservar la objetividad.

Este procedimiento tiene algunas ventajas formales, tales como que el proceso de estimación de los cuatro descriptores combinados se puede realizar directamente, si bien los estimadores tienden a infraestimar los valores reales. El problema que plantea es que debe asegurar que se disponga de los efectivos suficientes en las casillas marginales (véase fórmulas de cálculo de intervalos de confianza), para que se puedan realizar estimaciones precisas de los descriptores.

Por tanto, si la población diana es de bajo riesgo (baja prevalencia) se precisará un tamaño muestral muy importante para conseguir casillas marginales lo suficientemente pobladas. Los valores marginales previsibles para una muestra definida pueden ser calculados, si tenemos información sobre la prevalencia de enfermedad y la proporción de tests positivos, obtenida de otros estudios o de estudios piloto. En líneas generales, para una precisión en la sensibilidad y especificidad del 10% se necesitará un n de 25, para 5% de 100 y para 1% de 2.500.

Esta estrategia tiene indudables ventajas de validez porque definida la población diana y realizado el muestreo la relación con el espectro de pacientes es directa. Sin embargo, con esta estrategia todos los pacientes del estudio deberán tener completado el test y el diagnóstico, lo que puede resultar muy caro si se considera el tamaño muestral requerido. En otros casos, realizar algunos tests puede suponer riesgos excesivos para las personas.

Un aspecto a destacar (general para todas las estrategias) es la conveniencia de incluir los casos en los que el resultado del test (o más raramente del estándar) es dudoso. Evitar esta precaución plantea un problema de validez que se denomina *sesgo por exclusión de indeterminados*[29,43]. Si se han producido resultados indeterminados (x, y) no sabemos bien dónde hay que asignarlos, pero en las columnas verticales dispondremos en realidad de tres casillas: (a, x, c) en un caso y (b, y, c) en el otro. Podemos calcular la sensibilidad como a/a + b, pero realmente deberíamos calcular a/a + x + c, luego estamos sobrestimando la Se. De modo análogo podemos calcular la especificidad como d/b + d pero realmente el cálculo debía ser d/b + y + d. En consecuencia, la exclusión de indeterminados produce generalmente un error sistemático con sobrestimación de los índices citados.

«Estrategia retrospectiva» (a partir del diagnóstico) (fig. 3). En ella se parte de una muestra de la población diana de tamaño T a la que se le aplica el diagnóstico. Obtenemos dos subpoblaciones de pacientes, unos con enfermedad (D+) y otros sin enfermedad (D–). De cada una de esas subpoblaciones se toma una muestra representativa de sujetos con enfermedad (N1) y de sujetos sin enfermedad (N2). A ambas muestras (N1 + N2) se les aplica el test en estudio. Podría invocarse una cierta analogía de arquitectura con los estudios de casos y controles en el sentido de que la enfermedad clasifica a los sujetos y el test se explora a continuación.

El proceso de estimación de los descriptores es directo para la sensibilidad y especificidad, puesto que los valores verticales provienen directamente de la población (fig. 2). Para los valores predictivos los resultados deben estimarse a partir de los estimadores bayesianos (trabajan con probabilidades condicionales), puesto que los valores de N1 y N2 no representan, respectivamente, a la prevalencia de enfermedad y su complementario.

Esta estrategia es estadísticamente más potente que la estrategia simultánea (tanto más cuanto más se aleje

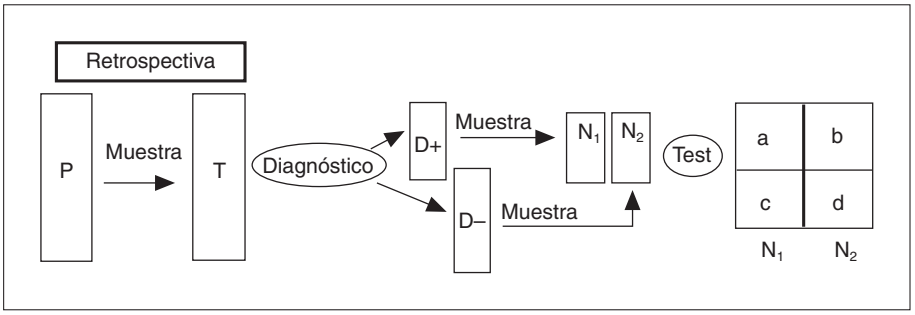

Fig. 4. Estrategias de muestreo en los estudios de evaluación de tests diagnósticos: retrospectiva; P: población de estudio; M: muestra primera. D+ = enfermos; D– = no enfermos; N_1 y N_2: muestras segundas.

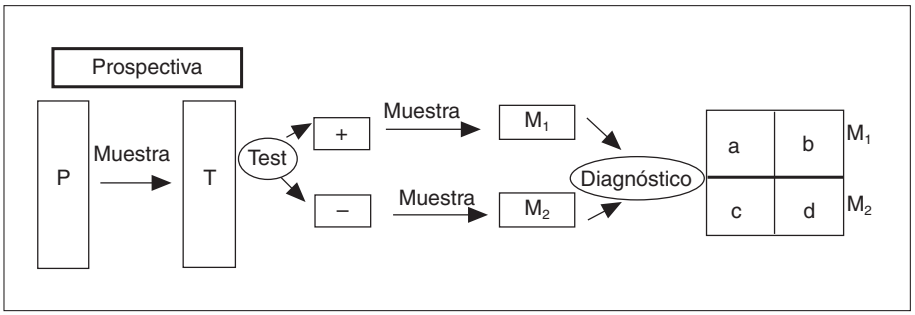

Fig. 5. Estrategias de muestreo en los estudios de evaluación de tests diagnósticos: prospectiva; P: población de estudio; M: muestra primera M_1 y M_2 = muestras segundas.

de una prevalencia 0,50 de enfermedad), y adquiere su máxima potencia cuando las muestras N1 y N2 son iguales (N1 = N2). Sin embargo, debe asegurarse que existan al menos 10 casos en la casilla marginal menos poblada. Otra ventaja de este diseño, respecto de la estrategia simultánea, es que convierte al estudio en muy coste-efectivo, si el coste del diagnóstico es bajo y el del test alto. Por tanto, sería un diseño a considerar cuando se trata de evaluar exploraciones de alta tecnología.

Desde el punto de vista bioético, un aspecto a destacar de este diseño es que la realización del test es totalmente superflua para el paciente, que ya está bien clasificado según el estándar. Por ello (y con carácter general para otros diseños) los riesgos del test deben ser cuidadosamente valorados para respetar el principio de *no maleficencia*, y el consentimiento informado debe ser obtenido, en respeto del principio de *autonomía*.

Estrategia prospectiva (a partir del test) (fig. 4). En ella se parte de una muestra de la población diana del estudio de tamaño N. Cada uno de los sujetos recibe el test, y se obtienen dos subpoblaciones de pacientes: con test positivo y con test negativo, respectivamente. A continuación se toma una muestra representativa de los pacientes con test positivo (M1) y una muestra (M2) de los sujetos con test negativo. Ambos grupos (M1 + M2) reciben a continuación el diagnóstico.

El proceso de estimación es directo para los valores predictivos positivo y negativo, y debe estimarse indirectamente la sensibilidad y especificidad a partir de estimadores bayesianos.

Esta estrategia es, también, más potente que la simultánea (tanto más cuanto más se aleje de 0,50 la proporción de tests positivos), y adquiere su potencia máxima si las muestras M1 y M2 son iguales (M1 = M2), y debe asegurarse al menos 10 casos en la casilla marginal menos poblada.

Un atractivo de esta estrategia es que se parece al proceso habitual en la actividad clínica, en efecto en la clínica se realiza primero el test y a continuación se realiza el *gold standard*, pero si comporta riesgos, sólo se aplicará el *gold standard* si el resultado del test es positivo. Este proceder puede plantear problemas en los estudios de diagnóstico, es el llamado *sesgo de verificación diagnóstica*[36,37,44]. Por ejemplo, imaginemos un estudio realizado a partir de la actividad clínica, en el que se compara el test de esfuerzo con la coronariografía. La probabilidad de indicar y realizar coronariografía a los sujetos con test positivo es mayor que a los sujetos con test negativo (actuar con más celo investigador plantearía problemas bioéticos). La probabilidad de entrar en el estudio (fracción de muestreo) es superior en el grupo de test positivo que en el de test negativo.

En los cálculos esto supone que las casillas a y b correspondientes a los tests positivos estarán artificialmente más pobladas y por ello la sensibilidad (a/a + b) estará sobrestimada y la especificidad (d/b + d) infraestimada.

Frecuentemente, como en nuestro ejemplo, es imposible evitar este sesgo, una solución puede ser desplegar un esfuerzo activo de seguimiento clínico de los episodios que ocurren en los pacientes con test negativo, de modo que podamos en un futuro reclasificarlos como enfermos o no enfermos, aunque no tengan coronariografía. Una segunda alternativa, si se dispone de otras informaciones, es la corrección matemática de esos resultados[45].

Existe una cuarta estrategia denominada seudorretrospectiva y que consiste en tomar una muestra N1 de una población de alto riesgo de enfermedad y una muestra N2 de una población de bajo riesgo de enfermedad. Ambas muestras reciben diagnóstico y test y los resultados se combinan como si se tratase de una estrategia prospectiva. Por las razones señaladas, a propósito del espectro de enfermedad, esta estrategia produce estimadores sesgados de sensibilidad y especificidad.

Esta estrategia seudorretrospectiva no es válida en sentido estricto. Algunos autores proponen su uso para las fases iniciales de evaluación de un test, si bien dentro de un proceso general de evaluación y planteando una analogía con las fases de desarrollo de fármacos (o fases del ensayo clínico). Así, Feinstein[29] propone una fase I en la que el test se probaría en grupos de sujetos claramente enfermos y claramente sa-

nos (estrategia seudorretrospectiva), una fase II con diseño retrospectivo en la que se ampliaría el espectro de la comparación, una fase III con diseño retrospectivo en la que se identificarían los tipos clínicos y patológicos de enfermedad y los grupos de comorbilidad y una fase IV con diseño simultáneo.

CONCLUSIÓN

Los estudios de evaluación de tests diagnósticos proporcionan conocimiento imprescindible para el uso de la probabilidad en el diagnóstico. Sin embargo, para que ese conocimiento sea válido y preciso el diseño de estos estudios debe respetar, en lo posible, una serie de condiciones que se han señalado a lo largo del artículo. En consecuencia cabe señalar como conclusiones:

En primer lugar, en contra de la frase con la que se iniciaba el artículo, en la que Laín Entralgo[46] describía la medicina del XIX, la práctica clínica es hoy día bastante más que la combinación de patología y *gramática parda*. En efecto, la incorporación de tecnologías y de modelos de reflexión al diagnóstico ha cambiado sustancialmente el panorama. En segundo lugar, la realización de estudios de evaluación de tests diagnósticos requiere un profundo conocimiento de las patologías, un conocimiento de los posibles errores a evitar y un diseño específico que optimice el esfuerzo investigador. En tercer lugar, para la incorporación del conocimiento sobre tests diagnósticos a la práctica clínica a partir de la bibliografía científica, es preciso disponer de una capacidad de juicio que permita una lectura crítica de la bibliografía publicada[47-49].

AGRADECIMIENTO

Los autores agradecen a V. Abraira, J. Gómez, J.J. Mira, A. Burls, J. Latour, M. Esteban y V. Mainar sus interesantes comentarios a versiones previas de este manuscrito. Asimismo, agradecen a Alicia Picó su inestimable ayuda en la edición del manuscrito. Procede el eximente habitual.

BIBLIOGRAFÍA

1. Sox HC, Blatt MA, Higgins MC, Marton KI. Differential diagnosis. En: Sox HC, Blatt MA, Higgins MC, Marton KI, editores. Medical Decision Making. Boston: Butterworths 1988; 2: 9-26.
2. Tversky A, Kahneman D. Judgement under uncertainty: heuristic and biases. Science 1974; 185: 1.124-1.131.
3. Diamon GA, Forrester JS. An epistemological model of clinical judgment. Am J Med 1983; 75: 129-137.
4. Sackett DL, Haynes Rb, Guyatt GH, Tugwell P. Interpretación de los datos diagnósticos. En: Sackett DL, Haynes RB, Guyatt GH, Tugwell P, Sackett DL, Haynes RB et al, editores. Epidemiología clínica. Ciencia básica para la medicina clínica (2.ª ed.). Buenos Aires: Editorial Médica Panamericana, 1994; 34-61.
5. Pauker SG, Kasirer JP. The threshold approach to clinical decision making. N Engl J Med 1980; 302: 1.109-1.111.
6. Pauker SG, Kasirer JP. Decision Analysis. N Engl J Med 1987; 316: 250-257.
7. Rifkin RO, Hood WB. Bayesian analysis of electrocardiographic stress testing. N Engl J Med 1977; 297: 681-686.
8. Diamond GA, Forrester JF. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. N Engl J Med 1979; 300: 1.350-1.358.
9. Wartofsky MW. La medida. En: Wartofsky MW, editor. Introducción a la filosofía de la ciencia. Madrid: Alianza Universidad, 1987; 7: 204-239.
10. Latour J, Abraira V, Cabello JB, López J. Las mediciones clínicas en cardiología: validez y errores de medición. Rev Esp Cardiol 1997; 50: 117-129.
11. Hlatky MA, Mark DB, Califf RM, Pryor DB. Angina, myocardial ischemia and coronary disease: Gold standards, operational definitions and correlations. J Clin Epidemiol 1989; 42: 381-384.
12. Trask N, Califf RM, Conley MJ, Kong Y, Peter R, Lee KL et al. Accuracy and interobserver variability of coronary cineangiography: A comparison with postmortem evaluation. J Am Coll Cardiol 1984; 3: 1.145-1.154.
13. Sanz ML, Mancini J, LeFree MT, Mickelson JK, Starling MR, Vogel RA et al. Variability of quantitative digital substraction coronary angiography before and after percutaneous transluminal coronary angioplasty. Am J Cardiol 1987; 60: 55-60.
14. Narula J, Chopra P, Talwar KR, Reddy KS, Vasan RS, Tandon R et al. Does endomyocardial biopsy aid in the diagnosis of active rheumatic carditis? Circulation 1993; 88: 2.198-2.205.
15. Shiffman RN. Guideline maintenance and revision. 50 years of the Jones criteria for diagnosis of Rheumatic fever. Arch Pediatr Adolesc Med 1995; 149: 727-732.
16. Joseph L, Gyorkos TW. Inferences for likelihood ratios in the absence of a gold standard. Med Decis Making 1996; 16: 412-417.
17. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-171.
18. Phelps CE, Hutson A. Estimating Diagnostic tests accuracy using a Fuzy gold standard. Med Decis Making 1995; 15: 44-57.
19. Rose GA. Chest pain questionnaire. Milbank Mem Fund Q 1965; 43: 32-39.
20. Sorlie PD, Coper L, Shreiner PJ, Rosamond W, Szklo M. Repeatability and validity of the Rose questionaire for angina pectoris in the atheroesclerosis risk in communities study. J Clin Epidemiol 1996; 49: 719-725.
21. Williams MJ, Marwick TH, O'Gorman D, Foale RA. Comparison of exercise echocardiography with an exercise score to diagnose Coronary Artery Disease in women. Am J Cardiol 1994; 74: 435-438.
22. Feinstein AR. Basic principles in the structure of clinimetric indexes. En: Feinstein AR, editor. Clinimetrics. New Haven, CT: Yale University Press, 1987; 23-43.
23. McNeil BJ, Keeler E, Adelstein SJ, Prime on certain elements of medical decision making. N Engl J Med 1975; 293: 211-215.
24. Detrano R, Alcedo E, Passalaqua M, Friis R. Exercise electrocardiographic variables: a critical appraisal. J Am Coll Cardiol 1986; 4: 86-87.
25. Ilsley C, Canepa-Anson R, Westgate C, Webb S, Richards A, Poole-Wilson P. Influence of R wave analysis upon diagnostic accuracy of exercise testing in women. Br Heart J 1982; 48: 161-168.
26. Kligfield P, Okin PM. Evolution of the exercise electrocardiogram. Am J Cardiol 1994; 73: 1: 209-1.210.
27. Robert AR, Melin JA, Detry JM. Logistic discriminant analysis improves diagnostic accuracy of exercise testing for coronary artery disease in women. Circulation 1991; 83: 1.202-1.209.
28. Iliceto S, Galiuto L, Marangelli V, Rizzon P. Clinical use of stress echocardiography: factors affecting diagnostic accuracy. Eur Heart J 1994; 15: 672-680.
29. Feinstein AR. Diagnostic and spectral markers. En: Feinstein AR, editor. Clinical Epidemiology. The architecture of clinical research. Filadelfia: Saunders Co., 1985; 597-631.

30. Pozo F. La eficacia de las pruebas diagnósticas (I). Med Clin (Barc) 1988; 90: 779-785.
31. Pozo F. La eficacia de las pruebas diagnósticas (II). Med Clin (Barc) 1988; 91: 177-183.
32. Milton JS, Tsokos JO. Inferencias sobre proporciones. En: Milton JS, Tsokos JO, editores. Estadística para biología y ciencias de la salud. Madrid: Interamericana. McGraw-Hill, 1987; 225-248.
33. Coughlin SS, Picle LW. Sensitivity and specificity-like measures of the validity of a Diagnostic test that are corrected for chance agreement. Epidemiology 1992; 3: 178-181.
34. Brenner H, Gellefer O. Chance corrected measures of the validity of a binary Diagnostic test. J Clin Epidemiol 1994; 47: 627-633.
35. Fleiss JL. The measurement of interrater agreement. En: Fleiss JL, editor. Statistical methods for rates and proportions (2.ª ed.). Toronto: Willey, 1981; 212-236.
36. Ransohof DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299: 926-930.
37. Begg CB. Biases in the assessment of diagnostic tests. Statistic in Medicine 1987; 6: 411-423.
38. Lachs ML, Nachamin I, Edelstein PH, Goldman J, Feinstein AR. Spectrum bias in the evaluation of diagnosis test: lessons from the rapid dipstick test of urinary tract infection. Ann Intern Med 1992; 117: 135-140.
39. Knotterus JA, Leffers P. The influence of referal patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992; 45: 1.143-1.154.
40. Garber CE, Carleton RA, Heller GV. Comparison of «rose questionaire angina» to exercise thallium scintigraphy: different findings in males and females. J Clin Epidemiol 1992; 715-720.
41. Blass EB, Follansbee WP, Orchard TJ. Comparison of supplemented Rose questionnaire to exercise thallium testing in men and women. J Clin Epidemiol 1989; 42: 385-393.
42. Van der Schouw YT, Van duk R, Verveek ALM. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic test. J Clin Epidemiol 1995; 48: 417-422.
43. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic test. J Chron Dis 1986; 39: 575-584.
44. Choi BCK. Sensitivity and specificity of a single Diagnostic test in the presence of work-up bias. J Clin Epidemiol 1992; 45: 581-586.
45. Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigren IR et al. The importance of work-up (verification) bias correction in assessing The accuracy of SPECT Thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol 1996; 49: 735-742.
46. Laín Entralgo P. La relación médico-enfermo. Madrid: Alianza Editorial, 1983; 215.
47. Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine working group. User's guides to medical literature VI. How to use an article about diagnostic tests. A. Are the results of the study valid? JAMA 1994; 271: 389-391.
48. Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine working group. User's guides to medical literature VI. How to use an article about diagnostic test. B. What the results and will they help me in caring for my patients? JAMA 1994; 271: 703-707.
49. Sackett DL, Richardson WS, Rosemberg W, Haynes RB. Evidence-Based Medicine. How to practice and teach EBM. Nueva York: Churchill-Livingstone, 1977; 81-89, 118-128.