



Estrategias diagnósticas en epidemiología clínica

- **Enrique Ardila MD, Profesor Asistente. Departamento de Medicina Interna, Universidad Nacional de Colombia. Hospital San Juan de Dios.**
- **Gilles Landrivon, René Ecochard, Francois Delahaye, Cyrille Colin eYves Matillon. Reseau International d'Epidemiologie Clinique Francophone, RECIF. Centre d'Information Médicale des Hospices Civils de Lyon.**

JUSTIFICACION DE LA EXISTENCIA DE TECNICAS EVALUATIVAS DE LOS DIFERENTES PROCEDIMIENTOS DIAGNOSTICOS

La rápida evolución de los conocimientos médicos en los últimos 30 años, tanto en el aspecto diagnóstico como terapéutico, ha hecho más difícil la toma de decisiones en el campo de la salud. Además, se han planteado interrogantes por parte de las personas teóricamente beneficiadas, acerca de su eficacia e inocuidad. En la práctica diaria actual, los médicos utilizan tecnologías avanzadas y enfoques científicos muy elaborados. Estas nuevas técnicas médicas, a menudo demasiado costosas, no siempre sustituyen las clásicas. El desarrollo de estas tecnologías podría llevar a plantear ciertos cuestionamientos de tipo ético sobre su utilidad para mejorar la salud y la posibilidad de que la población pudiera disfrutar de ellas.

Nuestra formación médica inicial ha estado centrada en el enfoque analítico del diagnóstico positivo y diferencial; es decir, el diagnóstico preciso y exhaustivo de todas las alteraciones que podrían afectar a nuestros enfermos. Este tipo de orientación obliga a los médicos a recolectar la mayor cantidad de información posible, mediante los exámenes de laboratorio y de imagenología, con el fin de llegar a un diagnóstico certero, descartando diagnósticos diferenciales. Una vez que se ha establecido el diagnóstico se emplean tecnologías y procedimientos avanzados para resolver el problema a nuestro paciente. Algunos procedimientos se utilizan antes de que su verdadera eficacia haya sido probada, basados en la simple hipótesis de que la nueva tecnología pueda ser más útil que la anterior.

Estas nuevas técnicas diagnósticas consumen una parte importante del presupuesto de salud, debido a su proliferación, utilización frecuente y gran complejidad. El problema radica en establecer la verdadera influencia que ejercen sobre la calidad del cuidado médico.

Esta evaluación abarca tres aspectos importantes en relación al paciente, el médico y la sociedad.

Con relación al paciente: su principal preocupación, cuando consulta al médico, es mejorar su estado de salud. Por lo tanto, la cualidad más importante de una prueba diagnóstica es su eficacia, en el sentido de utilidad y su inocuidad.

En la determinación de la eficacia de una prueba intervienen sus características técnicas y su precisión. La capacidad técnica es la adecuada relación entre los resultados suministrados y los objetivos que de ella se esperan.

La precisión o el grado de exactitud, ataña directamente al paciente (y por supuesto a su médico). Cuando un paciente consulta porque súbitamente comienza a presentar disnea, que se agrava rápidamente, y su médico sospecha que pudiera tratarse de un cuadro clínico de embolismo pulmonar, el enfermo deberá ser sometido a una serie de exámenes diagnósticos. Iniciando por los más sencillos (radiografía de tórax y electrocardiograma, determinación de gases arteriales), para continuar luego, avanzando en complejidad, con gammagrafía pulmonar (ventilación-perfusión), y si el caso lo requiere, una angiografía pulmonar por substracción digital o la determinación, por la técnica de ELISA, de los niveles de D-dímero.

El diagnóstico evocado con estos exámenes podría llegar a causar confusión y angustia, tanto al médico, como al enfermo, especialmente cuando los resultados de los análisis son positivos, y el paciente no presenta la entidad que estamos sospechando. Este sería el caso de los falsos positivos.

También se puede tropezar con el problema de que a pesar de continuar con la sospecha clínica de embolismo pulmonar, los análisis no sean confirmatorios del diagnóstico y más tarde encontramos que efectivamente nuestro paciente sí tenía un

embolismo pulmonar y no le suministramos el tratamiento adecuado. Este sería el caso de los falsos negativos.

La capacidad para identificar aquellos verdaderos positivos o sensibilidad y la capacidad para identificar verdaderos negativos o especificidad, definen la precisión de un test.

Inocuidad. Es importante para el enfermo, que el examen sea realizado en las mejores condiciones de confort y seguridad. Si efectuar un examen diagnóstico resulta desagradable, doloroso o peligroso, deberá replantearse la verdadera necesidad de su realización en términos del valor de la información que nos pudiera suministrar. Así mismo, el interés, la calidad y la importancia de éste en términos de su capacidad para modificar la conducta médica. En nuestro ejemplo, este sería el caso de la angiografía pulmonar.

Con relación al médico: frecuentemente, es a solicitud del médico que el paciente se somete a la realización de un procedimiento diagnóstico, pero en ocasiones, el mismo paciente podría estar interesado y solicitarlo; en ambos casos existe el mismo interés sobre su eficacia e inocuidad.

La existencia de una patología determinada se sospecha en función de los síntomas con los cuales cursa y de la frecuencia con que ésta se presenta dentro de un grupo determinado. Llamamos a esto probabilidad pre-test de la enfermedad, pero su confirmación sólo será puesta en evidencia después de haber realizado el examen en el cual tenemos más confianza y muy seguramente mayor experiencia. A esto lo llamamos probabilidad post-test y podría seguramente cambiar la actitud y la decisión médica.

Un examen positivo debe comprometer al médico a emprender un tratamiento, si el existe, para la enfermedad diagnosticada. El resultado deberá ser lo suficientemente positivo para que se justifique iniciar una terapéutica. Al contrario, frente a un test negativo, ¿hasta dónde el médico puede estar seguro de que su paciente estaría libre de la patología investigada?

En ese momento el médico se podría preguntar ¿hasta dónde es evidente que su paciente presenta un cuadro de embolismo pulmonar, frente a una gammagrafía pulmonar llamada positiva?

En otras palabras, ¿cuál sería el valor predictivo positivo y cuál el valor predictivo negativo de un examen?, y ¿cómo se podría establecer el umbral de un test?

Con relación a la sociedad: en este punto confluyen los argumentos concernientes al paciente y al médico. La inocuidad es un parámetro que incumbe a la sociedad, en la medida en que el eventual daño derivado de la práctica de procedimientos diagnósticos podría convertirse en carga social. Desde el mismo

punto de vista, las posibles repercusiones económicas juegan también un papel importante en la selección de un examen. Por estas razones, estaría justificado efectuar estudios de economía de la salud para confirmar la verdadera eficacia de una prueba diagnóstica.

Validez de un método diagnóstico. Se define como la capacidad para identificar tanto a los sujetos que padecen la enfermedad, como a quienes presentan otra patología. Establecer un diagnóstico es un procedimiento imperfecto que puede manifestarse en términos de probabilidad, más que en términos de certeza. Un examen diagnóstico, tendrá sentido solamente si su resultado es capaz de modificar la probabilidad de la enfermedad. Es decir, reducir la probabilidad a un nivel tal que el tratamiento no sea llevado a cabo, o de aumentarla hasta un punto en el cual el tratamiento se justifique.

Probabilidad pre-prueba (pre-test). Antes de prescribir la realización de una gammagrafía pulmonar, el médico deberá tener un alto grado de sospecha clínica de que su paciente pueda presentar un embolismo pulmonar.

Al referirse a un paciente en particular, el término a utilizar sería “probabilidad pre-prueba”. En caso de que se trate de un grupo, utilizaremos el término prevalencia. La prevalencia de una entidad determinada varía de acuerdo a si los sujetos de una población se encuentran en una de las tres situaciones siguientes: a quienes se les practica un examen de tamizaje, quienes consultan a su médico por la aparición de síntomas concretos, y quienes son remitidos a la consulta especializada, luego de ser vistos por el médico general.

Probabilidad post-prueba. Esta es la probabilidad más importante en clínica. Correspondría a la pregunta “¿Qué significa el resultado de un examen diagnóstico?”

Después de llevar a cabo un examen especializado, la probabilidad de padecer una enfermedad depende de tres elementos: del grupo al cual se pertenezca, de los tres enumerados anteriormente, de las características del examen (cantidad de falsos positivos y negativos) y de la positividad o la negatividad de la prueba.

Sensibilidad y especificidad de una prueba. Para presentar los resultados de un examen, lo más simple es ilustrar las diferentes situaciones posibles en una tabla de 2x2.

La gammagrafía pulmonar puede ser considerada como positiva (anormal) o negativa (normal) y la enfermedad estar presente o ausente. Esto llevará a cuatro posibles interpretaciones en los resultados del test: dos son correctas y dos son falsas. La prueba da una respuesta justa cuando la respuesta es positiva en presencia de la patología investigada (verdaderos positivos VP) o negativa

Tabla 1. Sensibilidad y especificidad de una prueba.

		EMBOLISMO PULMONAR	
		Presente	Ausente
Gamagrafia	+	V P	F P
	-	F N	V N

VP= Verdadero positivo; VN= Verdadero negativo
FP= Falso positivo; FN= Falso negativo

en ausencia de ésta (verdaderos negativos VN). Es falsa cuando aparece como positiva en ausencia de la enfermedad (falsos positivos FP) o negativa en presencia de la enfermedad (falsos negativos FN). Podemos por lo tanto definir:

Sensibilidad: probabilidad de tener una prueba positiva cuando se padece la enfermedad: cantidad de verdaderos positivos.

Especificidad: probabilidad de tener una prueba negativa, cuando no se padece la enfermedad: cantidad de verdaderos negativos. También la cantidad de falsos positivos y falsos negativos.

La proporción de resultados correctos, positivos o negativos, definen la exactitud (o precisión) de una prueba. El resultado del examen va a cambiar la opinión del médico en cuanto al estado probable del paciente, si el examen es confiable.

Si la prueba es negativa y sabemos que puede haber pocos falsos negativos (prueba sensible) el médico puede pensar, con alta probabilidad, que su paciente no padece la enfermedad. En el caso de los gases arteriales (sensibilidad del 95% y especificidad del 50%) ante la sospecha de un embolismo pulmonar y resultado normal (negativos): no se justificará ordenar más exámenes diagnósticos, no se iniciará ningún tratamiento, se suspenderá el tratamiento, si éste se ha instaurado.

Si el test es positivo y sabemos que puede haber pocos falsos positivos (prueba específica), como es el caso de la angiografía pulmonar cuando se sospecha embolismo pulmonar, el médico estaría autorizado para iniciar el tratamiento adecuado.

El resultado del examen, si éste no es confiable, no cambiará la opinión del médico sobre el diagnóstico de su paciente.

Si se sabe que la prueba tiene un gran número de falsos negativos, (radiografía de tórax en el caso de sospecha de embolismo pulmonar, en el cual los signos radiológicos pueden no estar o desaparecer rápidamente) el médico continuará con su diagnóstico clínico. En general el médico clínico conoce la sensibilidad y la

especificidad (exactitud) de los exámenes que está acostumbrado a solicitar. El problema que se plantea al médico es saber si su paciente tiene o no la enfermedad después del resultado de la prueba. En este punto se deben definir:

El valor predictivo positivo de la prueba: probabilidad de padecer la enfermedad cuando el examen es positivo (caso de la angiografía pulmonar cuando se sospecha embolismo pulmonar).

El valor predictivo negativo de la prueba: probabilidad de no tener la enfermedad cuando el resultado es negativo (caso de los gases arteriales cuando se sospecha embolismo pulmonar).

Esto nos lleva al concepto de probabilidad post-prueba.

Relación entre probabilidad pre y post-prueba. Existe un sistema matemático para estimar si el juicio que estamos efectuando, frente a una enfermedad dada, es correcto (valores predictivos del test) en función de la validez del test. Si conocemos la prevalencia de la enfermedad en la población y la sensibilidad y especificidad de una prueba, es posible, gracias al teorema de Bayes, determinar la probabilidad particular de que un sujeto con un test positivo realmente padezca la enfermedad que estamos estudiando, y que un sujeto negativo, esté realmente libre de la misma.

Tabla 2. Relación entre probabilidad pre y post-prueba.

VALOR PREDICTIVO =	$\frac{(PREVALENCIA)(SENSIBILIDAD)}{(PREVALENCIA)(SENSIBILIDAD) + (1-PREVALENCIA)(1-ESPECIFICIDAD)}$
VALOR PREDICTIVO =	$\frac{(1-PREVALENCIA)(ESPECIFICIDAD)}{(1-PREVALENCIA)(ESPECIFICIDAD) + (PREVALENCIA)(1-SENSIBILIDAD)}$

Para los valores predictivos de un test, la prevalencia de la enfermedad es un factor más importante que la sensibilidad y la especificidad. En efecto, encontramos que la sensibilidad y la especificidad varían raramente más de dos veces, mientras que la prevalencia puede variar sobre una escala más grande.

No suscita el mismo grado de sospecha de embolismo pulmonar un hombre joven deportista, que consulta por haber presentado súbitamente dificultad respiratoria, que una paciente obesa de mediana edad a quien se le ha practicado recientemente una cirugía pélvica. Evidentemente, la posibilidad de presentar un embolismo pulmonar es muy remota, en el primer caso, mientras que en el segundo ésta es mucho más alta.

Umbral de tratamiento. Frecuentemente al médico debe intervenir, o al contrario abstenerse de hacerlo si no posee una gran certeza. Esta situación no es rara en el caso de sospecha de embolismo pulmonar, debido a que el único examen con una buena sensibilidad y especificidad es la angiografía pulmonar, y ésta no se realiza en todos los hospitales.

Para efectos prácticos, el médico deberá estar en capacidad de evaluar clínicamente un umbral de probabilidad en virtud del cual valdría la pena someter al paciente a tratamiento. Este umbral es llamado umbral de tratamiento.

Si la diferencia entre la probabilidad pre-prueba y el umbral de tratamiento es grande, solamente un examen de muy buena calidad (pocos falsos positivos y pocos falsos negativos, según el caso) podrán cambiar la decisión terapéutica. Si el examen posee esas cualidades, vale la pena efectuarlo; de lo contrario pierde su interés diagnóstico.

Compromiso entre sensibilidad especificidad. Es evidente que frente a un caso en el cual se sospecha embolismo pulmonar, el médico quisiera disponer de una prueba altamente sensible y altamente específica. Deberá encontrarse una relación entre la sensibilidad y la especificidad.

La curva ROC (Receiver Operating Characteristic Curve) es un medio de expresar esta relación, permitiendo describir la precisión de una prueba, y podría ser utilizada para comparar de manera práctica dos exámenes diferentes en el diagnóstico de una misma enfermedad.

Una prueba podría ser considerada como normal (negativa), o bien como anormal (positiva). Entre una prueba normal y una anormal, se puede definir toda una escala de probabilidades de padecer la enfermedad; por ejemplo: imposible, incierto, improbable, poco probable, poco dudoso, ligeramente dudoso, dudoso, posible, muy probable, probable, manifiesto y certero. En la práctica, se podría reducir a una escala continua: sospecha mínima, ligera, poca o grande.

En la escogencia del punto que marca el límite entre la normalidad y la anormalidad, para un mismo test, radica en la selección del valor del umbral de positividad del test, por debajo del cual se considera negativo y como positivo por encima del mismo. Esta decisión se toma en forma arbitraria. Pero es esta escogencia la que va a determinar la indicación de un examen más especializado, que puede estar basado en los diferentes grados de sospecha que nos ha aportado el primer test.

Si nos apoyamos en una sospecha mínima para ordenar un nuevo examen, es decir, si fijamos un umbral de positividad demasiado bajo, muy pocos pacientes que realmente tengan la enfermedad se pasarán por alto, pero un gran número de sujetos que no la padecen serán considerados como enfermos.

Una prueba de este tipo es un examen muy sensible, pero poco específico y presenta un número elevado de falsos positivos. En el otro extremo, si nos apoyamos sólo en una gran sospecha de la enfermedad para ordenar un examen más especializado; es decir utilizamos un umbral de positividad elevado, la mayoría de

pacientes que presenten una primera prueba positiva, tendrán realmente esta patología, pero un número grande pacientes serán excluidos, aún teniendo la enfermedad, y falsamente serán llamados sanos. En este caso el test propuesto, es un examen muy específico, que sacrifica sensibilidad, dando lugar a un número elevado de falsos negativos. La relación entre la sensibilidad, lugar a un número elevado de falsos negativos. La relación entre la sensibilidad y la especificidad se puede expresar construyendo la curva ROC. Esta representa el equilibrio entre la cantidad de diagnósticos positivos entre los enfermos (cantidad de verdaderos positivos o sensibilidad) y cantidad de diagnósticos positivos en los sanos (cantidad de falsos positivos, o falta de especificidad, o 1-especificidad).

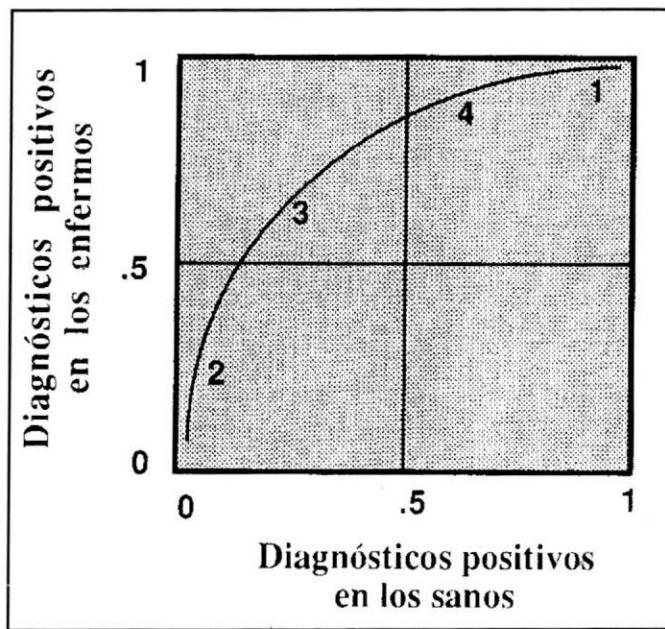


Figura 1. Representación de las variaciones conjuntas de la cantidad de falsos positivos en función del umbral de positividad seleccionado: cuatro escogencias del umbral, cuatro consecuencias diferentes.

La prueba más sensible considerada como positivo si poco probable y negativo si improbable. Con él se obtiene una gran sensibilidad, pero a costa de un gran número de falsos positivos (punto 1 de la curva)

La prueba menos sensible sólo considera los resultados manifiestos como positivos y hace necesario llevar a cabo un examen más especializado. Los resultados probables permanecen como negativos haciendo su sensibilidad muy débil, pero sin falsos positivos (punto 2 de la curva).

Dos situaciones intermedias (puntos 3 y 4 de la curva).

Punto 1: examen clínico. Es un examen de sensibilidad mediocre y mala especificidad.

Punto 2: gammagrafía pulmonar. Sensibilidad de 98% y

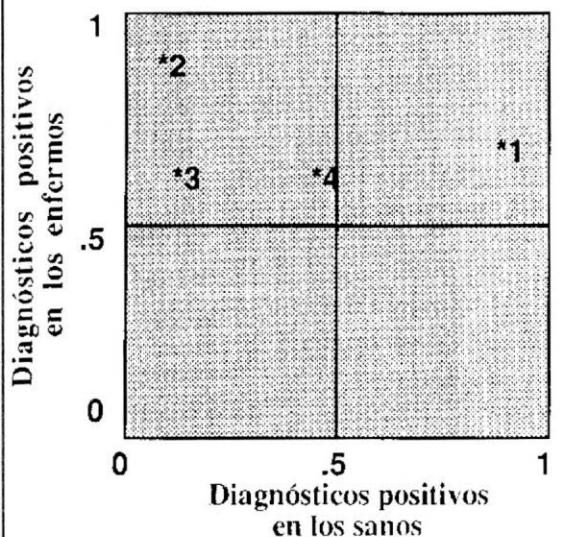


Figura 2. Tres diferentes tipos de exámenes paraclínicos utilizados en el diagnóstico de embolismo pulmonar.

especificidad del 90%.

Punto 3: corresponde a un examen de sensibilidad mediocre y buena especificidad como podría ser la radiografía de tórax

Punto 4: corresponde a un examen de sensibilidad mala y especificidad mediocre como sería el caso de un electrocardiograma.

¿El examen menos explorador? presenta una curva ROCv que cae cerca de la diagonal del gráfico.

Entre más grande sea la superficie bajo la curva, mejor es el test y mejor es el compromiso entre sensibilidad y especificidad.

Vimos que podemos situar los diferentes exámenes diagnósticos en tres zonas particulares del gráfico. La diagonal corresponde a curvas de test "nulos": encontrando el mejor compromiso para una sensibilidad de 50% que equivale a una situación al azar.

Determinación del umbral de positividad de una prueba.

Regresamos en definitiva al problema de definir el límite entre normalidad y anormalidad. Es la determinación del valor del umbral la que define la positividad y la negatividad de un test. Esto depende de la importancia que le demos al riesgo de tener resultados falsos.

La selección del umbral depende de dos grandes factores: probabilidad pre-prueba, que variará según el tipo de examen y el contexto en el que él se efectúe. Por ejemplo: tamizaje, consulta general, consulta especializada; y equilibrio costo\beneficio, es decir el equilibrio entre: de una parte, el inconveniente de tratar a una persona sana (costo), y de otra parte, la ventaja de tratar un enfermo (beneficio).

Si los falsos positivos y los falsos negativos tienen los mismos costos (este término se refiere a costos monetarios o humanos), será natural seleccionar como umbral, o como examen diagnóstico en el que nos pueda suministrar la mayor parte de diagnósticos correctos.

Así, se podría maximizar la relación:

$$\frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{conjunto de pacientes}}$$

Lo contrario ocurriría cuando el costo de los falsos negativos es más importante.

Reproductibilidad. Se denomina reproductibilidad a la proporción de medidas que son idénticas. La cuantificación en la reproductibilidad de un test se basa en el cálculo del coeficiente Kappa.

La concordancia entre los resultados de las medidas efectuadas por diferentes observadores puede ser más o menos elevada. Ello depende de las cualidades reales del test (sensibilidad y especificidad), como también de la interpretación. El coeficiente Kappa evalúa el nivel de concordancia alcanzado en relación con aquél debido simplemente al azar: El grado de acuerdo que se obtiene, en relación con el acuerdo máximo que se puede obtener sólo por el azar.

O= Acuerdo observado.

E= Acuerdo debido sólo al azar.

Kappa= $(O-E)/(1-E)$.

Kappa= 1 para una concordancia perfecta.

Kappa= 0 cuando no hay más concordancia que la debida al azar. Valores negativos de Kappa son posibles, indicando que hay menos acuerdo que el que puede ser observado debido al azar.

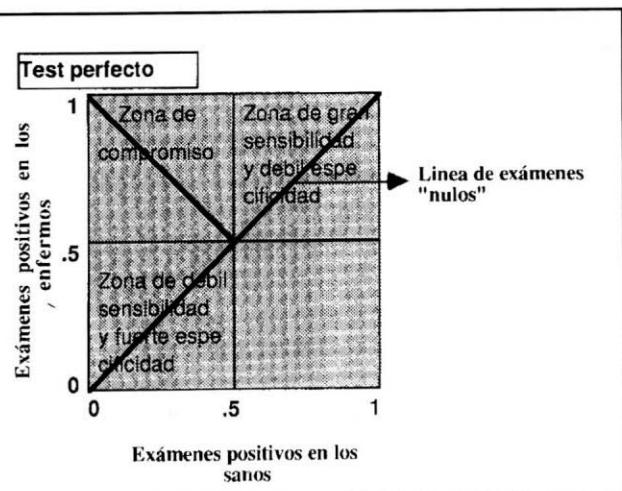


Figura 3. Resumen de las diferentes situaciones encontradas.

solamente.

Exámenes múltiples. En la patología pulmonar no se puede afirmar o descartar fácilmente la existencia de un tromboembolismo, después de la utilización de un solo examen diagnóstico. En efecto, los exámenes perfectos (en lo ideal 100% de sensibilidad y de especificidad) son muy raros. En ocasiones nos hallamos en una situación, en la cual la probabilidad de la enfermedad, después de realizar un examen, no se modificará de manera decisiva. Sería necesario entonces llevar a cabo una terapéutica, o nuevos tests diagnósticos.

En este último caso la estrategia diagnóstica estará constituida por varios exámenes múltiples, para lo cual hay dos formas de proceder: en paralelo, es decir todos los exámenes al mismo tiempo, o bien, en serie, o sea, unos después de otros. El grado de urgencia es lo que diferencia los dos enfoques diagnósticos. Las pruebas utilizados en paralelo: por ejemplo examen clínico, radiografía de tórax, electrocardiograma, gammagrafía, angiografía pulmonar. Este enfoque se realiza (todos los exámenes solicitados al mismo tiempo), cuando se desea obtener diagnóstico rápidamente, como en el caso de situaciones médicas quirúrgicas de urgencia o en los pacientes vistos ambulatoriamente. La sensibilidad, y por tanto el valor predictivo negativo, es mejor para el conjunto de estos exámenes solicitados en paralelo, que cuando se hace para uno separadamente. Lo contrario ocurre para los exámenes en serie. Así, después de una secuencia de tests, es menos probable pasar por alto una enfermedad, pero los diagnósticos falsos positivos serán más frecuentes.

Estos exámenes en paralelo son particularmente útiles cuando el médico necesita un examen muy sensible, pero tiene a su disposición solamente dos o tres pruebas moderadamente sensibles. Utilizar exámenes en paralelo resulta ser una estrategia diagnóstica de mejor sensibilidad, al costo de los tratamientos inútiles originados en los falsos positivos.

Las pruebas utilizados en serie: por ejemplo cuando el examen clínico es sospechoso. Este enfoque se lleva a cabo cuando no es indispensable obtener los resultados rápidamente, como en la práctica de la consulta, en oposición a las situaciones hospitalarias de urgencia. Los tests en serie son utilizados igualmente, cuando algunos de ellos son costosos o presentan algún riesgo para el paciente: tal test en efecto no se emplea sino en el caso de que no exista uno más simple o menos agresivo, que nos sirva para investigar la entidad que estamos sospechando. Esta estrategia, aunque consume menos recursos económicos que los exámenes en paralelo (debido a que la utilización de un segundo examen dependerá del resultado del primero, que puede ser inútil) sí requiere más tiempo.

Las pruebas utilizadas en serie aumentan la especificidad y el valor predictivo positivo de la estrategia diagnóstica, pero disminuyen la sensibilidad y el valor predictivo negativo. El

médico y el paciente van a estar más seguros de que la enfermedad realmente existe cuando el examen es positivo, reduciendo el riesgo de falsos negativos. Este enfoque en serie es particularmente útil cuando ninguna de las pruebas disponibles es muy específico. En tal caso la prueba más específica deberá ser utilizado inicialmente; el más sensible vendrá después. De esta forma, aunque a menos pacientes se le practicarán los mismos exámenes, el mismo número de casos será diagnosticado.

Evaluación de las estrategias diagnósticas

Es más fácil evaluar un tratamiento médico, que una prueba diagnóstica. Una intervención terapéutica es en efecto evaluada típicamente mediante un ensayo clínico característico que comprende dos grupos, distribuidos al azar: uno que recibe el tratamiento en cuestión y otro el placebo. Al final del estudio, se podrá estar seguro de que existe una diferencia entre los dos grupos (por el seguimiento, tasa de mortalidad o nivel sanguíneo de algún elemento característico) que pueda ser atribuida solamente a la intervención practicada. No ocurre lo mismo en la evaluación de una estrategia diagnóstica, que debe realizarse a diferentes niveles.

¿Qué se deberá tener en cuenta?

La sensibilidad y la especificidad del test: estos son los criterios clásicos de la efectividad de un test y deberán intervenir en la escogencia de un examen diagnóstico. Pero, ¿entre un test con pocos falsos negativos y otro con falsos pocos positivos, cuál escoger? Vimos que la curva ROC puede aportar elementos de decisión, haciendo evidente el compromiso entre las diferentes cualidades de un test.

El impacto diagnóstico: este puede ser evaluado respondiendo diferentes preguntas:

¿El resultado de la prueba va a modificar la secuencia del procedimiento diagnóstico o va a evitar la realización de otros exámenes más invasivos?

¿La prueba utilizada es capaz de reemplazar otras pruebas?

¿El resultado del test reduce la probabilidad de tener la enfermedad hasta el punto que podemos considerar que el tratamiento no es necesario?; en tal caso, ¿el paciente podría estar seguro de que el riesgo de falsos negativos no es muy elevado?

Impacto terapéutico: el principio fundamental radica en el hecho de que su utilización pueda modificar la espongencia terapéutica en el paciente a quien se le practicó el test.

Impacto sobre el resultado clínico: finalmente, para el paciente el punto importante no es el diagnóstico, que no es más que una etapa, sino el tratamiento. Llegar a comprobar un diagnóstico para el cual no existe tratamiento, no es precisamente lo que más le interesa al paciente.

Los datos necesarios en la evaluación son con frecuencia

insuficientes o están ausentes?

Una prueba diagnóstica dada puede ser preferida para los pacientes que presentan *a priori* un riesgo bajo (probabilidad pretest o prevalencia) de la enfermedad, mientras que otro será preferido para los que presentan una probabilidad *a priori* elevada.

Ya vimos la relación entre la prevalencia y el valor predictivo de un test. Los estudios sobre las pruebas diagnósticas raramente involucran este tipo de información.

La mayoría de los pacientes involucrados en esos estudios, incluyendo los ensayos clínicos, frecuentemente constituyen una pequeña proporción de aquellos a quienes realmente se les debería practicar el estudio.

Efectivamente, son excluidos de estos estudios los pacientes que presentan más de un problema patológico.

Igualmente, es difícil tener una idea precisa, cuando se presentan resultados negativos, dada la falta de información sobre la cantidad de falsos negativos.

El estudio clínico aleatorizado en el análisis de una prueba diagnóstica.

Se dice que un estudio clínico aleatorizado bien concebido y bien ejecutado, es la mejor forma de comparar el poder de dos tecnologías. Sin embargo, cuando se evalúan dos estrategias diagnósticas, esto tiene sus límites.

El costo. Este puede ser elevado, especialmente cuando se desea apreciar las cualidades de un nuevo test en la forma de aplicación ideal, debido a que en ese caso es necesario estandarizar los cuidados del paciente.

Población estudiada. Se corre el riesgo de que ésta sea insuficiente, especialmente cuando se trata de enfermedades raras o de enfermedades crónicas de evolución lenta.

El factor tiempo. Puede ser también un obstáculo en la evaluación de una tecnología diagnóstica. En efecto, un período de tiempo largo puede ser necesario en el caso de que el test requiera una metodología especial. Tomando en cuenta que la tecnología evoluciona continuamente, el estudio puede llegar a ser obsoleto antes de que sus conclusiones sean formuladas.

Los sesgos de selección. La tecnología médica es evaluada, la mayoría de las veces, en los hospitales universitarios en donde los pacientes, las indicaciones, la patología y la experiencia no suelen ser representativas del conjunto de la población médica, por lo que no siempre es posible generalizar los resultados.

La concepción de un ensayo clínico

Es difícil, por razones éticas, comparar un grupo que se beneficiaría de un examen diagnóstico, con un grupo control al cual no se le practicará, a menos que estuviéramos convencidos, por argumentos clínicos contundentes, de que los sujetos del grupo no están enfermos. En ese último caso, sin embargo, la información suministrada por el estudio representaría muy poco interés.

Si escogemos un ensayo clínico como medio para el estudio de una nueva tecnología diagnóstica, nos podemos imaginar dos concepciones diferentes, según se considere o no el estándar de referencia.

En primer lugar, si el interés del estudio de la prueba radica en la consecuencias clínicas para el paciente, no es necesario hacer referencia a un estándar bien definido. En ese caso el ensayo viene bien para comparar dos pruebas diferentes: todo sujeto que necesite de la prueba puede ser incluido dentro del estudio, el cual será capaz de definir cuál es la mejor. No obstante, este tipo de estudio no puede suministrar información sobre la cantidad de verdaderos positivos y de falsos positivos, debido a la falta de un estándar de referencia (*gold standard*), y por lo tanto no se podrá definir la precisión de la misma.

En segundo lugar, se podría disponer de un estándar de referencia con el cual se evaluarán todos los pacientes incluidos en el estudio. Así es posible determinar la precisión de la prueba estudiada, la cual se podrá comparar con una ya conocida y se podrá medir el efecto sobre la morbilidad o la decisión terapéutica. Sin embargo, el hecho de exigir un estándar de referencia va a excluir del estudio un cierto número de sujetos que son incapaces o no desean realizar exámenes suplementarios. Esto haría difícil la generalización de las conclusiones.

En otros términos, una prueba podría ser evaluada en dos situaciones diferentes. Su eficacia puede ser estudiada en condiciones ordinarias de utilización: todos los pacientes a quienes se les puede practicar el test son incluidos. Esto sería efectuado por el personal médico con cierto grado de experiencia y de habilidad. Este enfoque es sobre todo el del médico clínico "en su campo". Su eficacia puede ser estudiada dentro de las condiciones ideales de utilización. Pero si se quiere conocer la verdadera precisión de la prueba, tenemos que reducir las fuentes de variabilidad: pacientes muy seleccionados, pacientes que presentan exámenes en las mismas condiciones de preparación, utilización del personal más experimentado. Este último enfoque es sobre todo el del científico que participa de la actualización de una nueva tecnología diagnóstica. Pero la información que se obtiene en condiciones ideales de utilización reduce la posibilidad de generalización (o de aplicabilidad). Esta situación podría ser resumida en la Figura 4.

Problema particular en la interpretación de una prueba:

Experiencia del investigador: la habilidad del clínico puede evolucionar en el curso de un estudio de evaluación de exámenes diagnósticos. Siendo nueva la prueba, su experiencia al principio del estudio es limitada. Por consiguiente, se corre el riesgo de que la cantidad de resultados falsos (falsos positivos o falsos negativos) sea más elevada con relación al estándar de referencia a la iniciación del estudio que a su terminación.

Homogeneidad en la interpretación: lo ideal sería contar con el

mismo observador para interpretar todos los exámenes y realizar todas las medidas según los criterios de normalidad o de anormalidad estandarizados. Si varios observadores están implicados en el estudio, deberán tener todos la misma experiencia y la misma habilidad.

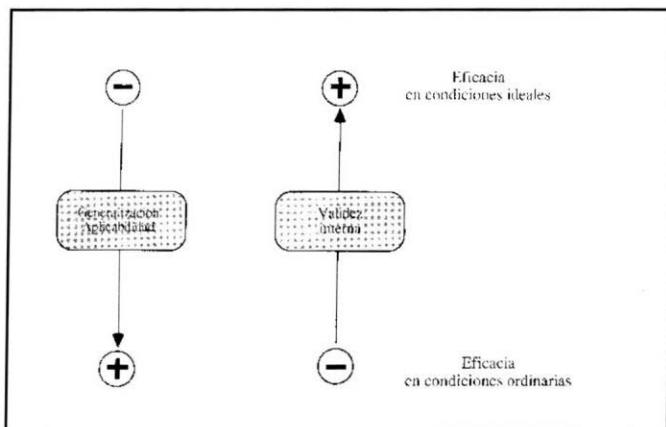


Figura 4. Eficacia de una prueba.

Objetividad: es absolutamente indispensable que la interpretación del test, objeto del estudio, sea independiente de la interpretación del test de referencia: el médico que interpreta la nueva prueba no deberá conocer el resultado del test de referencia. El médico que interprete el examen de referencia, no deberá conocer el resultado de la nueva prueba. Esto podría llevar a problemas éticos en la medida en que la decisión terapéutica, orientada a partir de los resultados, estará basada en una interpretación que no aportará un beneficio a todos los datos de otras pruebas disponibles.

Esquema general de un estudio de evaluación de exámenes diagnósticos

Selección de la situación clínica estudiada: frecuencia de la enfermedad que se supone la prueba va a diagnosticar, costo de la tenología, impacto del diagnóstico y acceso al estándar de referencia.

Definición de la población de pacientes a los cuales se les va a realizar la prueba: seleccionar los enfermos de tal forma que un gran espectro de la enfermedad sea representada (fases tempranas y fases avanzadas de la enfermedad) y excluir a los enfermos que presenten más de una patología.

Descripción de la técnica diagnóstica estudiada, así como el estándar de referencia, si existe y es utilizado, y definición del tipo de comparación; nueva tecnología que reemplaza la antigua; prueba A versus prueba B, nueva tecnología combinada con la antigua, comparando a la antigua sola: prueba A sólo, versus prueba A más B.

Definición del error tipo I (riesgo de afirmar que existe una

diferencia entre el nuevo y el antiguo test, cuando en la realidad no la hay) y del error tipo II (riesgo de afirmar que no hay diferencia entre el test nuevo y el antiguo, cuando en realidad sí la hay). Determinación del número de sujetos que serían sometidos al nuevo test diagnóstico estudiado, con el fin de detectar una diferencia de los parámetros estudiados entre el nuevo test y el antiguo.

Descripción del procedimiento de aleatorización, si se trata de un ensayo clínico, con sus complicaciones éticas.

Presentación de los parámetros a estudiar y de los análisis estadísticos implicados: Los valores *a priori* y *a posteriori* para establecer los valores predictivos, la curva ROC a propoada para evaluar los test cuyos resultados son presentados en forma de variables continuas. Esta es utilizada para comparar la precisión (validez) de dos pruebas, calculando las dos superficies situadas bajo la curva. La cantidad de falsos positivos y de falsos negativos, necesitan un estándar de referencia.

Descripción del proceso “doble ciego”, es decir el método por el cual el médico interpreta un nuevo examen, del cual no conoce los resultados de la prueba de referencia y viceversa.

Resultado del seguimiento de los pacientes (debido a que la razón por la cual se analiza un test diagnóstico es el impacto clínico).

Estudio de la reproductibilidad del test (coeficiente Kappa).

Descripción de los efectos secundarios del test.

Eventualmente, análisis económico.

Lectura crítica de la literatura médica con referencia a los exámenes diagnósticos.

El estudio de la lectura crítica de la literatura médica, referente a los exámenes diagnósticos, juzga la influencia que ésta pudiera tener en la práctica del lector.

El objetivo es apreciar la credibilidad de los resultados y definir la aplicabilidad en su propia práctica. Para que ésto se lleve a cabo, se podrían plantear ocho preguntas, para las cuales se deberá buscar la respuesta en la publicación. Una respuesta satisfactoria a cada una de ellas, garantizaría la calidad del estudio referente a los exámenes diagnósticos y su aplicabilidad en la práctica corriente.

¿La prueba puede ser comparada con un estándar de referencia?
¿La población de enfermos abarca un espectro importante de casos?

¿El sitio del estudio y la selección de los pacientes están descritos de una manera adecuada?

¿La validez de la prueba (sensibilidad, especificidad, valores predictivos) y la reproductibilidad de los resultados y/o la interpretación de los mismos han sido discutidas?

¿El término “normal” ha sido bien definido?

¿El examen es uno de los elementos de estrategia? ¿La estrategia

es evaluada dentro del contexto clínico habitual?

¿La forma de llevar a cabo el examen está descrita? (con el objeto de poder reproducir la técnica).

¿La utilidad real del examen ha sido determinada? (Gracias a su aplicación los enfermos se han mejorado).

AGRADECIMIENTOS

Los autores expresan su gratitud al doctor Francois Chapuis por su asesoría en la elaboración de los gráficos y al doctor Juan Manuel Arteaga Díaz por la lectura y corrección de la versión castellana del texto.

BIBLIOGRAFIA

1. Schwartz JS. Evaluating diagnostic test: What is done - What needs to be done. *Journal of General Int Med* 1984; 1: 266-267.
2. Sheps SB and Schechter MT. The assessment of diagnostic tests: A survey of current medical research *JAMA* 1984 1984; 252: 2418-2422.
3. Kent DL, Larson EB. Diagnostic technology assessment: Problems and prospects. *Ann Intern Med* 1988; 108: 759-761.
4. Jenicek M, Cleroux R. Epidemiologie. Principes, techniques et Applications. Paris Quebec. Editions Maloine. S-A 1993.
5. Sox HC. Probability theory in the use of diagnostic test: An introduction to critical study of the literature. *Ann Intern Med* 1986; 104: 60-66.
6. Fletcher RH, Fletcher SW, Wagner O. Clinical Epidemiology, The essentials. Baltimore: Williams and Wilkins 1988.
7. Philbrick JT, Horwitz, Feinstein AR. The limited spectrum of patients studies in exercise test research: analysing the tip of the iceberg *JAMA* 1987; 248: 2467-2470.
8. Annas GJ. Your money of your life: "Dumping" uninsured patients from hospital emergency wards. *American Journal of Public Health* 1986; 76: 74-77
9. Sox HC, Margulies I, Sox CH. Psychologically mediated effects of diagnostic tests. *Ann Intern Med* 1981; 95: 680-685.
10. Ransohoff DF, Ferstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic test. *N Engl J Med* 1978; 299: 926-930.
- 11.. Adelstein SJ. Pitfalls and bias in evaluating diagnostic technologies. In: Mac Neil BJ, Cravalho EG eds. Critical Issues in Medical Technology. Boston: Auburn House Publishing Company 1982: 67-79.
12. Sox H, Stern S, Owens D, Abrams HL. Assessment of Diagnostic Technology in Health Care Rationale, Methods, Problems and Directions. Washington D.C: Institute of Medicine National Academy Press 1989.
13. Freiman JA, Chalmers JC, Smith H, Kuebler RR. The importance of B, the type II error and sample size in the design and interpretation of the randomized clinical trial. *N Engl J Med* 1978; 299: 690-694.
14. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. Boston: Little, Brown and Company 1985.
15. Fowles FG, Fulton PM. Critical Appraisal of Published Research: introductory Guidelines. *Br Med J* 1991; 302: 1136-1140.
16. Quinn DA, Thompson T, Terrin ML. A prospective investigation of Guidelines. *Br Med J* 1991; 302: 1136-1140.
17. Matchar DB, Simel DL, Geweke JF, Feussner JR. A bayesian method for evaluating medical test operating characteristics when some patients' conditions fail to be diagnosed by the reference standard. *Med Decis Making* 1990; 10: 102-11.
18. Demers C, Ginsberg JS, Johnston M. D-dimer and thrombin-antithrombin III complexes in patients with clinically suspected pulmonary embolism. *Thromb-Haemost* 1992; 67: 408-412.
19. Bounameaux H. Significance of the determination of D-dimers in venous thrombo-embolic disease. *J Mal Vasc* 1992; 17 (Suppl. B): 91-92
20. Bounameaux H, Schneider PA, Slosman D. Plasma D-dimer in suspected pulmonary embolism: a comparison with pulmonary angiography and ventilation-perfusion-scintigraphy. *Blood Coagulation Fibrinolysis*. 1990; 1: 577-579.