

Error de medición o la capacidad operativa de las pruebas clínicas

Objetivo:

- Introducir sobre la metodología para la evaluación de las pruebas clínicas o diagnósticas
- Presentar la guía para la evaluación de un artículo sobre pruebas diagnósticas

Introducción:

Este capítulo exige quizás más que otros, una lectura atenta y detallada. El tema es sencillo, pero puede resultar confuso si no se van comprendiendo progresivamente los conceptos expuestos. Por ello, un buen consejo: lea despacio, piense en el sentido cada frase, relea todas las veces que le parezca necesario. Solamente avance en la medida que comprendió bien lo precedente. A la tarea !!!.

Debemos aceptar que el error diagnóstico es un hecho posible. Por lo tanto, las decisiones deben ser adoptadas hacia el tipo de error más tolerable para el paciente (*Fornety JA y col. The development of an index of high-risk pregnancy. Am J Obstet Gynecol 1982, 143: 501*). Es decir, aquel que menos afecte su seguridad, tanto en la ejecución misma de la prueba como en sus consecuencias para el manejo del caso.

Se afirma que “una prueba será aceptable cuando resulte ANORMAL en casi todos los individuos que presenten el daño y NORMAL en casi todos aquellos que no lo presenten” (*Griner PF y cols. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med. 1981: 94(4 Pt 2):557-92*). Entonces, para utilizar cabalmente una prueba, se debería conocer el tamaño relativo de los “casi”, pues utilizar pruebas con error de medición desconocido pueden poner en riesgo al paciente dado que inducen a la toma de decisiones equivocadas.

Se observa que el médico es más proclive a modificar su opinión sobre el estado del paciente si la prueba clínica solicitada presume que existe el daño buscado que si informa que no lo hay. Por otro lado, mayor cantidad de información no es sinónimo de mejores decisiones médicas, pero sí de mayores costos, molestias o riesgos. Las mejores decisiones están vinculadas a la calidad de la información y no a su cantidad.

¿Por qué decimos esto?. Porque habitualmente buscamos abundar en información sobre el estado del paciente solicitando exámenes complementarios de los cuales desconocemos su capacidad para detectar aquello que estamos buscando.

Haga la prueba:

¿de cuántas de las pruebas clínicas que usted utiliza a diario en la búsqueda de un diagnóstico conoce la proporción en la que pueden arrojarle un resultado falso?

Cierto grado de error es intrínseco a toda medición, aún en las consideradas más confiables. Puede afirmarse que no hay pruebas perfectas, entendiendo por esto a que jamás arrojen un resultado falso. Conocer las capacidades de los métodos diagnósticos empleados es fundamental para la correcta interpretación de los resultados.

¿Para qué solicitamos una prueba complementaria?

La respuesta correcta es: para reducir la incertidumbre sobre el estado del paciente.

Cuando accedemos a un paciente por primera vez ignoramos su situación y nos abocamos a la tarea de efectuar un diagnóstico, que es la actividad distintiva del médico (*Editor's choice: Diagnosis, diagnosis, diagnosis. BMJ 2002; 324 (March 2)*). Sin un diagnóstico correcto no hay tratamiento efectivo. Para ello recurrimos a la inspección, al interrogatorio y al examen físico. En estos pasos habremos mirado, preguntado y relevado aquellos datos que la situación sugiera intentando reducir nuestra incertidumbre inicial sobre el estado del paciente. En cada uno de esos pasos habremos aplicado criterios que irán permitiéndonos descartar ciertas alternativas y asumir otras. Así vamos avanzando en la elaboración del diagnóstico. Finalizado el primer contacto con el paciente, surgen dos posibilidades: que lo observado haya resultado suficiente para elaborar un diagnóstico e indicar una conducta o que persista un grado tal de incertidumbre que obligue a ahondar la investigación, lo que habitualmente resolvemos solicitando exámenes complementarios. Si dispusiéramos de pruebas perfectas, no habría incertidumbre: el estado del paciente será el que diga la prueba. Como no las tenemos, debemos conocer el presunto grado de error con el cual las pruebas indican el estado del paciente, para actuar en consecuencia.

Atención, la expresión “pruebas clínicas” no sólo se refiere a las que se llevan a cabo en el laboratorio, la sala de radiología u otros, sino que incluye a todos los pasos del examen clínico.

Por ejemplo, ¿Cuál es el error de medición de la maniobra de palpación del polo inferior de bazo?; ¿Siempre que suponga palparlo, habrá esplenomegalia y siempre que no lo palpe, significa que no la hay?. ¿Cuál es la proporción en que se produce uno y otro error?. ¿Le parece que sería útil conocerla?.

Otra: ¿Cuán confiable es la respuesta obtenida al preguntar sobre una adicción?. ¿Un “no” la descarta y un “sí” la admite?. ¿Estima que la probabilidad de error es similar para ambas alternativas?.

Seamos conscientes que avanzamos hacia el diagnóstico trepando por una escalera cuyos escalones no son lo suficientemente firmes. En muy contadas ocasiones, podemos encontrarnos ante una característica que orienta a un diagnóstico exacto. Son las llamadas “patognomónicas”. Su detección hace innecesaria otra investigación.

El error de medición puede surgir desde:

- Cuestionarios (deben detectar lo que se pretende. Una pregunta puede ser muy clara para el que la realiza pero confusa para el que responde)
- Aparatos (no siempre disponemos de equipos adecuadamente calibrados o en condiciones. A propósito, ¿cuando fue la última vez que hizo calibrar su balanza?)
- Técnicas de aplicación (puede disponerse del mejor aparato, pero emplearlo mal. La medición de la presión arterial es uno de los actos cotidianos de la práctica asistencial que más deplorablemente se efectúan)
- Criterios de interpretación (el mismo informe puede ser interpretado como anormal por un médico y normal por otro. Es muy importante considerar las condiciones o contexto en el cual se obtiene el dato)

Contenidos:

Hemos presentado el problema. Ahora concentrémonos en cómo medir la capacidad operativa de las pruebas para dar a cada una su verdadero valor. Es decir, veamos como podemos estimar la magnitud de los “casi” mencionados en el primer párrafo.

Abordaremos los siguientes ítems:

- Sensibilidad
- Especificidad
- Valor pronostico positivo
- Valor pronostico negativo
- Índice de eficiencia pronostica
- Curvas ROC
- Concordancia
- Diseño de investigación
- Guía de lectura crítica

La magnitud de los indicadores surge de los estudios de investigación pertinentes (ver diseño), que informan la relación que se observó entre los resultados de la prueba y el estado real de los participantes en la investigación efectuada. Para informar y elaborar los resultados de la investigación, se recurre a la herramienta básica epidemiológica, la tabla de contingencia de 2 por 2. La misma, convencionalmente se conforma ubicando en las columnas las alternativas básicas referidas al daño: pacientes con daño o pacientes sin daño y en las filas las alternativas de resultado de la prueba en evaluación: normal o anormal.

Para efectuar la correcta evaluación es importante que definir los siguientes ítems:

- Punto final: es el daño buscado (ej.: diabetes).
- Patrón Oro (“gold-standard”): es el criterio o procedimiento que se toma en consideración para definir la existencia real del punto final (ejemplo: biopsia de piel para definir la existencia de un microangiopatía).
- Punto de corte: criterio con el cual se califica a la prueba como normal o anormal, sugiriendo la presencia o ausencia del daño (ejemplo: glucemia $\leq 120 >$ mg%)

Nota: Al mencionar “daño”, estamos haciendo referencia a cualquier situación adversa a la salud de un individuo que interese detectar. Al hablar de los resultados de las pruebas corrientemente se emplean como sinónimos las expresiones anormal o positivo y normal o negativo. Estas expresiones deben ajustarse acorde a la prueba que se emplee. Por ejemplo, en ciertos casos podría ser mejor hablar de reactiva y no reactiva, o de elevada y normal, etc.

Tabla básica de contingencia:

		Daño según Patrón Oro		
		Presente	Ausente	
Daño según la prueba	Presente	a	b	a + b
	Ausente	c	d	c + d
		a + c	b + d	N

Los casilleros identificados con las letras a, b, c, y d, representan respectivamente los casos en que:

- la prueba sugirió correctamente la existencia de daño: casillero a (verdaderos dañados ó verdaderos positivos)
- la prueba sugirió la existencia de daño pero éste no existía: casillero b (falsos dañados ó falsos positivos)
- la prueba sugirió la inexistencia de daño pero éste estaba presente: casillero c (falsos no dañados ó falsos negativos)
- la prueba sugirió correctamente la inexistencia del daño: casillero d (verdaderos no dañados ó verdaderos negativos)

La sumatoria de a+b, informa la cantidad total de casos en los cuales la prueba sugirió la presencia de daño.

La sumatoria de c+d, informa la cantidad total de casos en los cuales la prueba sugirió la ausencia de daño.

La sumatoria de a+c, informa la cantidad total de casos en los cuales el daño estaba presente.

La sumatoria de b+d, informa la cantidad total de casos en los cuales el daño estaba ausente.

En el casillero N se informa la cantidad total de casos incluidos en el estudio.

Bien, ahora disponemos de la información dispuesta en la forma apropiada para calcular la capacidad operativa de la prueba en análisis. Pero, ¿cuáles son los indicadores epidemiológicos para medirla?.

Indicadores de Capacidad Operativa
Sensibilidad
Especificidad

Se llama SENSIBILIDAD a la proporción en la cual una prueba arroja un resultado anormal cuando es aplicada a personas con el daño.

Para el cálculo de la sensibilidad, participan los datos de los siguientes casilleros:

		Daño según Patrón Oro		
		Presente	Ausente	
Daño según la prueba	Presente	a	b	a + b
	Ausente	c	d	c + d
		a + c	b + d	N

Sensibilidad es el nombre de la proporción resultante entre el denominador “total de dañados” (casillero a + c) y el numerador “los dañados identificados por la prueba” (casillero a). Se calcula utilizando exclusivamente valores que se ubican en casilleros de la columna numérica izquierda del núcleo de la tabla.

Sensibilidad: $a / (a+c)$

Utilizando el mismo denominador (a + c), y como numerador la cantidad de casos en los que la prueba no detectó daño pero lo había (casillero c), se construye una proporción que se denomina falsos negativos.

Falsos negativos: $c / (a + c)$

Se llama ESPECIFICIDAD a la proporción en la cual una prueba arroja un resultado normal al ser aplicada sobre quienes no padecen el daño.

En el cálculo de la especificidad, participan los datos de los siguientes casilleros:

		Daño según Patrón Oro		
		Presente	Ausente	
Presencia de daño según la prueba	Presente	a	b	a + b
	Ausente	c	d	c + d
		a + c	b + d	N

Especificidad es el nombre de la proporción resultante entre el denominador “total sin daño” (casillero b + d) y el numerador “los no dañados identificados por la prueba” (casillero d). Se calcula utilizando exclusivamente valores presentes en la columna numérica derecha del núcleo de la tabla.

$$\text{Especificidad: } d / (b+d)$$

Utilizando el mismo denominador (b+d) y como numerador a cantidad de casos en los que la prueba anunció daño pero no lo había (casillero b), se construye una proporción que se denomina falsos positivos.

$$\text{Falsos positivos: } b / (b + d)$$

Dado que sensibilidad y especificidad son técnicamente proporciones e informan una probabilidad, sus valores van de 0 a 1. Si deseamos expresarlas en forma porcentual, multiplicamos el valor anterior por 100, y por ende, sus valores se expresarán de 0 a 100. Así es lo mismo decir que la sensibilidad de una prueba es 0.78 o del 78%.

Los falsos resultados, negativo o positivo, son la proporción que representa el complemento de la sensibilidad y especificidad, respectivamente, a la unidad, ya sea 1 o 100. Por ejemplo: si la especificidad es 80% (ó 0.8), falsos positivos, será 20% (ó 0.2).

Como ocurre con toda información numérica probabilística, ella no está completa si no se informan sus límites ó intervalos de confianza (IC, habitualmente se calculan para una confiabilidad del 95% y lo escribiremos como IC95%). Por ejemplo, informaremos la sensibilidad de la prueba del supuesto del párrafo anterior diciendo: 0.78 (IC95% 0.55 a 0.89).

¿Cómo deben considerarse estas capacidades para la elección de una prueba?

Para descartar la existencia de un daño se requiere una prueba muy sensible.

Al ser muy sensible identificará a casi todos los individuos que posean el daño. Si en un individuo determinado se aplica una prueba muy sensible y su resultado sugiere que no hay daño (resultado negativo), éste puede descartarse con confianza, ya que si lo hubiera la prueba habría dado anormal. Pero si la prueba sugiere que el daño está presente, no podría asegurarse que el mismo existe ya que al ser muy sensible es posible que se torne anormal frente a personas sin daño (falsos positivos).

Para aceptar la existencia de un daño se requiere una prueba muy específica

Si es muy específica, identificará a casi todos los individuos que no posean el daño. Si en un individuo determinado se aplica una prueba muy específica y su resultado sugiere la existencia de daño (resultado positivo), éste puede asumirse con confianza. Si la prueba sugiere que no hay daño, no podría asegurarse que el mismo no exista ya que al ser muy específica hay notorias probabilidades de que personas con daño hayan arrojado resultado normal (falsos negativos).

Repasando:

La sensibilidad es la respuesta a la pregunta sobre cuánta es la probabilidad de que un paciente con el daño tenga una prueba anormal. La especificidad es la respuesta al preguntar cuánta es la probabilidad de que un paciente sin el daño obtenga una prueba normal. Sensibilidad y especificidad son conceptos que expresan la capacidad intrínseca de la prueba y resultan valores estables para cada una. Una prueba puede presentar alta solamente una de esas dos características. Es controvertido utilizarla en la práctica para aprovechar exclusivamente la capacidad destacada. Como se verá más adelante, ambas capacidades deben ser suficientes, aunque una puede serlo más que otra. Si ambas no tienen ese carácter se inducirá a muchos errores en la calificación de los pacientes (falsos positivos y falsos negativos).

Si bien sensibilidad y especificidad son dos propiedades independientes no deben analizarse por separado al momento de considerar la elección de una prueba.

Estos dos conceptos (sensibilidad y especificidad), ¿son las respuestas que el médico necesita al asistir un paciente individual?. No.

Volviendo a la tabla contingencia, ambos conceptos se responden utilizando una columna ya que el denominador se halla al pie de la misma. Si observamos la tabla con atención, vemos que para disponer de esos denominadores debemos

conocer quienes padecen el daño y quienes no. He aquí el problema, el médico los está buscando. Obviamente, no dispone de ese dato.

La pregunta que se hace el médico es ¿cuánta es la probabilidad de que el paciente tenga el daño si el resultado de la prueba es anormal y cuánta de que no lo posea si es normal?. Es decir, el médico mira el comportamiento de las pruebas utilizando los casilleros alineados en las filas de la tabla de contingencia (recuerde que para sensibilidad y especificidad, eran los de las columnas). Lo que él dispone es el resultado de la prueba solicitada y desea saber en que proporción ese resultado identifica adecuadamente el estado del paciente. A este concepto se lo denomina “valor pronóstico”.

Indicadores de Capacidad Pronostica
Valor Pronóstico Positivo
Valor Pronóstico Negativo

Nota: también llamados valores predictivos

Se llama Valor Pronóstico Positivo a la proporción de verdaderos dañados entre todos los que la prueba anuncia como tales. Es la capacidad de una prueba para identificar a quienes están dañados.

		Daño según el Patrón Oro		
		Presente	Ausente	
Daño según la prueba	Presente	a	b	a + b
	Ausente	c	d	c + d
		a + c	b + d	N

Valor pronostico positivo es el nombre de la proporción resultante entre el denominador “total de individuos con prueba anormal” (casillero a + b) y el numerador “individuos con daño y prueba anormal” (casillero a). Se calcula utilizando datos presentes en la primera fila numérica del núcleo de la tabla.

Valor pronostico positivo: $a / (a+b)$

Se llama Valor Pronóstico Negativo a la proporción de verdaderos no dañados entre todos los que la prueba anuncia como tales. Es la capacidad de una prueba para anunciar a quienes no están dañados.

		Daño según el Patrón Oro		
		Presente	Ausente	
Daño según la prueba	Presente	a	b	a + b
	Ausente	c	d	c + d
		a + c	b + d	N

Valor pronostico negativo es el nombre que se le da a la proporción resultante entre el denominador “total de individuos con prueba normal” (casillero c + d) y el numerador “individuos sin daño y con prueba normal” (casillero d). Se calcula utilizando exclusivamente valores presentes en la segunda fila numérica del núcleo de la tabla.

Valor pronostico negativo: $d / (c+d)$

A diferencia de lo que ocurre con la sensibilidad y especificidad, los valores pronósticos no son estables para cada prueba sino que varían acorde a la frecuencia del daño (prevalencia) en la población estudiada

(Prevalencia: $(a + c) / N$)

Por lo tanto, no pueden ser correctamente interpretados sin conocer la prevalencia del daño en la población a la que pertenece el paciente que estudiamos.

Vamos a realizar unos ejercicios para facilitar la interpretación de lo que hemos comentado. Supongamos que se efectuó una investigación que incluyó 100 individuos, con la finalidad de estudiar la capacidad operativa de la prueba XX. La tabla de contingencia final, resultó la siguiente:

		Daño según el Patrón Oro		
		Presente	Ausente	
Daño según la prueba XX	Presente	12	4	16
	Ausente	18	66	84
		30	70	100

Calcule:

- sensibilidad:
- especificidad:
- valor pronostico positivo:
- valor pronostico negativo:
- prevalencia:

Ahora, imagine que se llevó a cabo otra investigación similar para la misma prueba, pero en otra población de muy diferentes características y arrojó los siguientes datos:

		Daño según el Patrón Oro		
		Presente	Ausente	
Daño según la prueba XX	Presente	4	5	9
	Ausente	6	85	91
		10	90	100

Calcule:

- sensibilidad:
- especificidad:
- valor pronostico positivo:
- valor pronostico negativo:
- prevalencia:

¿Qué se modificó?

Sensibilidad y especificidad permanecieron estables, puesto que se trató de la misma prueba. (recuerde que habíamos mencionado que eran valores intrínsecos de la prueba y estables). La prevalencia se modificó, ya que era una población diferente de la anterior (30% en el primer caso y 10%, en el segundo).

Atención: Los valores predictivos también se modificaron notoriamente.

Por ejemplo, veamos la situación del valor pronostico positivo. En el primer caso la prueba le informaba al médico que un paciente con un resultado positivo tenía una probabilidad del 75% de padecer el daño. Esto condiciona fuertemente la conducta clínica a seguir. En el segundo caso, ante el mismo resultado la probabilidad de padecer el daño es del 45%. Es decir, que es menos probable que lo padezca que no lo padezca (45% vs. 65%). La conducta médica no puede ser la misma que en el caso anterior, donde era mucho más probable que el paciente padeciera el daño.

Dado que estos conceptos no están demasiado incorporados, el médico suele considerar a los resultados de las pruebas independientemente de la probabilidad de daño (prevalencia) que presente el paciente. Si la prueba fuera perfecta, no habría problemas, sus resultados serían siempre exactos sin importar la prevalencia del daño. Pero, no hay pruebas perfectas.

En un estudio realizado en Boston, EE.UU., en 1960, para identificar los valores de glucemia según el método de Somoyi-Nelson, a las 2 horas desde la sobrecarga, con el propósito de identificar a los diabéticos, se observó, que:

Corte mg%	Sensib	Especif	F -	F +
80	97.1	25.5	2.9	75.5
100	88.6	69.8	11.4	30.2
120	71.4	92.5	28.6	7.5
140	57.1	99.4	42.9	0.6
160	47.1	99.8	52.9	0.2
180	38.6	100	61.4	0

En esta tabla que presenta los resultados elaborados, podemos observar que el comportamiento de la prueba se modificó acorde al nivel de glucemia utilizado (es el punto de corte, que es el criterio con el cual se distingue un resultado anormal de uno normal). Probablemente, el nivel de 120 mg% es el que guarda mejor equilibrio entre las propiedades de la prueba. Observe que en la medida que el punto de corte se hace menos crítico (exigir glucemias menos elevadas para admitir diabetes), aumenta la sensibilidad. Expresado de otra manera, al disminuir la exigencia del punto de corte, cada vez menos personas con daños como el buscado no serán identificadas por la prueba. Si se aplica la prueba con este criterio, se seleccionarán como “con daño” a muchos individuos que no lo poseen, con lo cual la aplicación de la prueba se torna ineficiente. Veamos que ocurre si se admite un punto de corte muy exigente, por ej.: 180 mg%. Los individuos que la prueba identifique seguramente padecerán el daño pero esta seguridad será a costa de no identificar a muchos que padecen formas más leves del daño.

Cada vez que modificamos el punto de corte hacia menos exigencia, aumentará la sensibilidad en detrimento de la especificidad y vice-versa.

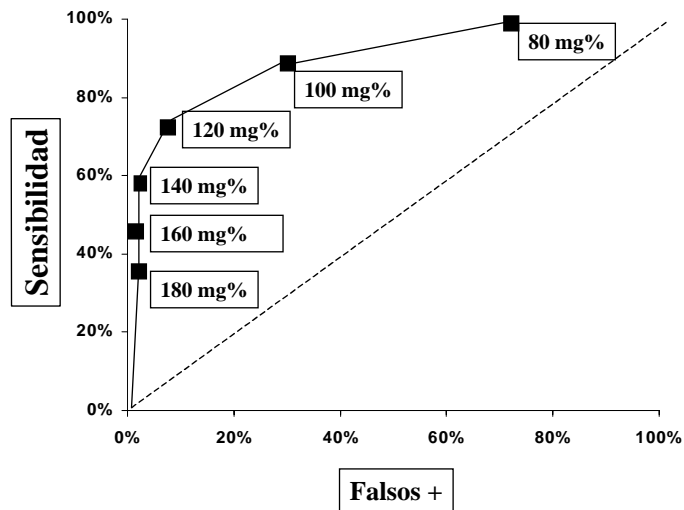
Este ejemplo nos permite concluir que uno de los elementos más influyente sobre las capacidades operativas de la pruebas es el punto de corte. Repetimos: criterio con el cual se distingue a un resultado anormal de uno normal.

¿Cómo seleccionar el punto de corte adecuado?

El PUNTO DE CORTE apropiado depende de:

- la gravedad del daño si no es tratado: un daño con consecuencias graves debería tratarse aún a costa de tratar a alguien no dañado.

- la factibilidad y eficacia del tratamiento: lo anterior debe estar en concordancia con que el tratamiento sea factible y eficaz. Si es muy costoso o sólo moderadamente eficaz, deberá reservarse para casos en los que no haya dudas diagnósticas.
- el costo y riesgos de los tratamientos aplicados empíricamente a los sin daño: si el tratamiento es de bajo costo y mínimos riesgos, puede admitirse su aplicación a individuos sin el daño, en aras de no dejar ningún dañado sin tratar.
- la prevalencia del daño: si el daño es muy prevalente hay que tratar de utilizar pruebas con alta especificidad.
- un mismo método puede aplicarse con diferentes puntos de corte, ajustados a cada finalidad.



Herramienta para identificar el punto de corte más apropiado

Son las llamadas CURVAS ROC (Receiver Operating Characteristic Curves).

Es una curva que expresa la relación existente entre los verdaderos positivos (sensibilidad) con los falsos positivos (inversa de la especificidad).

Se representa sobre un gráfico cartesiano y la curva expresa mayor poder discriminatorio de la prueba cuánto mayor sea el área bajo la curva.

Veamos el siguiente ejemplo:

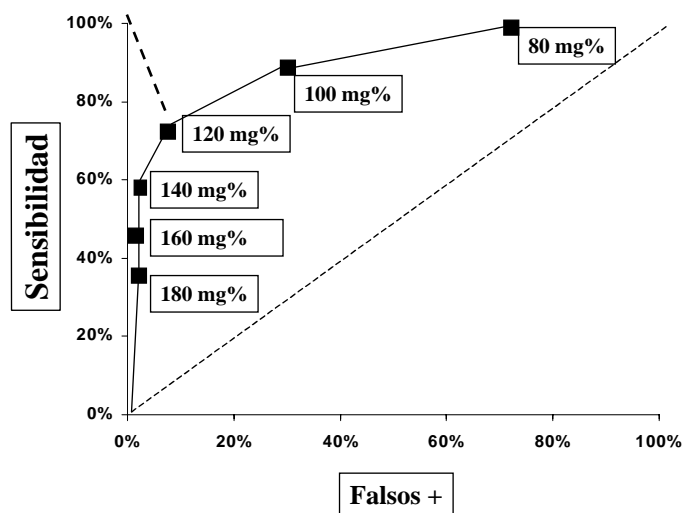
Esta curva está construida con los datos procedentes de la tabla anteriormente presentada sobre los valores de glucemia.

La línea punteada es la bisectriz del gráfico y representa la línea de no discriminación. Es decir, cualquier punto que se ubique en esa línea dice que la prueba con ese punto de corte produce igual proporción de resultados ciertos que

falsos. Con ello, asumimos que la prueba no posee ninguna utilidad práctica. Cuanto menos si el punto se ubica por debajo de esa línea.

El punto de corte ideal en una prueba perfecta sería aquel que se muestre con 100% de sensibilidad y 0% de falsos positivos. ¿Dónde se ubicaría este punto? Al tope del eje de las abscisas y al inicio del eje de las coordenadas.

Para el caso de la glucemia, ¿cuál es el punto de corte recomendable?. Será el que se ubique más próximo al punto ideal. En este caso, 120 mg%, tal como admitimos al analizar la tabla.



Si trazamos una línea que una ese punto con el ideal, veremos que es la más corta que puede trazarse, expresando que es el punto de corte estudiado que se halla más próximo al ideal.

Existen procedimientos estadísticos que hacen esa evaluación más precisa, pero entendemos que para el médico asistencial es suficiente con el método sugerido.

¿Cómo se construye una curva ROC?

Con los datos obtenidos en el estudio correspondiente, se construyen tablas de contingencia para cada uno de los puntos de corte que se desean probar. El total de casos (N) será el mismo para todas las tablas, al igual que la cantidad de individuos con el daño y sin él, dado que el criterio con el cual se diferencia un dañado de un no dañado no se modifica (recordemos, este es el llamado punto final). Lo que será diferente en cada tabla serán las cantidades en los casilleros centrales, ya que al modificar el punto de corte se modificará la cantidad de casos que la prueba calificó de una u otra condición (dañados o no dañados).

En cada una de esas tablas se calcularán la sensibilidad y los falsos positivos. Luego, esos valores se grafican sobre el gráfico pertinente.

APLICACION DE LAS PRUEBAS CLINICAS

La correcta asistencia de los pacientes depende un diagnóstico preciso.

El diagnóstico también depende de la calidad de las pruebas clínicas aplicadas y del conocimiento de su capacidad.

¿Cuándo aplicar pruebas complementarias de diagnóstico o pronóstico?

La respuesta es: cuando al finalizar el examen clínico, el médico persiste con incertidumbre respecto al verdadero estado del paciente.

Además han sido sugeridas las siguientes pautas (*Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med. 1981; 94(4 Pt 2):557-92*)

- Si el daño buscado es lo suficientemente frecuente cómo para justificar el esfuerzo de su detección.
- Si es capaz de producir consecuencias significativas.
- Si existe algún cuidado que revierta la evolución.
- Si la detección en estadios asintomáticos aporta claros beneficios sobre la detección en estadios sintomáticos precoces.

¿Qué le interesa al médico clínico?

Conocer el beneficio que otorgará la aplicación de una prueba determinada para reducir su incertidumbre sobre el estado del paciente. Por ende, para la práctica diaria lo importante es conocer la capacidad de la prueba indicada para predecir el estado o pronóstico del paciente.

Surge, entonces, una importante pregunta:

¿Con la aplicación de la prueba indicada, en qué medida reducirá el médico su desconocimiento sobre el estado y pronóstico del paciente?

Dado que no hay pruebas perfectas, para responderla hay que determinar:

- la probabilidad de daño al recibir la prueba
- la capacidad de la prueba para modificarla

Ya vimos que el indicador “valor pronostico” no es confiable, por su inestabilidad frente a las modificaciones de la prevalencia. Se requiere un indicador libre de estas influencias.

Para evaluar las pruebas clínicas con facilidad y confiabilidad, se requiere un instrumento que:

- Permita seleccionar el punto de corte que otorgue la mejor operatividad a cada prueba clínica.
- Resulte poco influenciado por la prevalencia.

Es instrumento es el llamado “INDICE DE EFICIENCIA PRONOSTICA” (IEP), “likelihood ratio”, en lengua inglesa. También se lo ha denominado “razón de

momios”, índice de verosimilitud”, etc. Expresa la capacidad o potencia de una prueba para modificar la probabilidad previa de daño asignada al paciente. Dado que se calcula a partir de la sensibilidad y especificidad, recibe escasa o nula influencia de la prevalencia de daño. Por lo tanto, es estable para cada prueba y punto de corte. De allí su importancia práctica. Se calculan un IEP positivo y uno negativo. El primero se vincula a la presencia de daño, mientras que el segundo a su ausencia.

Cálculo del IEP:

IEP(+): $\text{Sensibilidad} / (1 - \text{Especificidad})$

IEP(-): $(1 - \text{Sensibilidad}) / \text{Especificidad}$

¿Por qué la referencia al valor 1?

La prueba será más eficiente cuánto mayor a 1 sea el valor del IEP(+) y cuando menor a 1 sea el valor del IEP(-).

Si la sensibilidad resultara igual a (1 – especificidad), denominador y numerador de la ecuación serán iguales y el resultado será 1. Como este valor se va a utilizar para multiplicar la probabilidad previa de daño, al hacerlo por 1, aquel no se modificará. Por ende, la aplicación de la prueba no resulta de ninguna utilidad.

Los siguientes valores pueden ser útiles para orientar sobre la capacidad de una prueba:

Valores IEP	Capacidad
IEP(+) ≥ 10 ó IEP(-) ≤ 0.1	suficiente
IEP(+) $\geq 5 < 10$ ó IEP(-) $> 0.1 \leq 0.2$	moderada
IEP(+) $\geq 2 < 5$ ó IEP(-) $> 0.2 \leq 0.5$	escasa
IEP(+) $\geq 1 < 2$ ó IEP(-) $> 0.5 < 1$	insignificante

A esta altura de la lectura, intente escribir la definición de sensibilidad y de especificidad, sin volver atrás:

Sensibilidad:

Especificidad:

Aplicamos la fórmula para calcular el IEP y vemos que:

IEP + : $0.99 / (1 - 0.99) = 99$

IEP - : $(1 - 0.99) / 0.99 = 0.01$

Supongamos ahora que tenemos una prueba cuya sensibilidad es del 40% y su especificidad del 80%. Por ejemplo, en este rango se encuentran la mayoría de las pruebas para evaluación de la salud fetal, así como muchos de los procedimientos que los clínicos practican a diario.

Aplicamos la fórmula, y vemos que:

$$\text{IEP} + : 0.40 / (1 - 0.80) = 2$$

$$\text{IEP} - : (1 - 0.40) / 0.80 = 0.75$$

Observemos que diferentes son los valores del IEP en una y otra alternativa. En la primera, prueba con una capacidad operativa suficiente, el IEP presenta valores muchos más distantes de 1 que en el segundo caso, prueba con capacidad entre escasa e insignificante.

¿Recuerdan el ejercicio que realizamos para mostrar la variación del valor pronóstico ante la modificación de la prevalencia?. Volvamos al mismo y calculemos el IEP(+) para una prevalencia del 30% y del 10%.

Caso 1: Prueba con sensibilidad 40% y especificidad 95%; prevalencia del daño: 30%.

Calcule el IEP(+) =

Caso 2: Prueba con sensibilidad 40% y especificidad 95%; prevalencia del daño: 10%.

Calcule el IEP(+) =

¿Qué le dicen los resultados?

¿El IEP(+) quedó estable ante la modificación de la prevalencia?.

La respuesta es Sí. El IEP(+) es 8 en ambos casos, ya que no se modificaron ni la sensibilidad ni la especificidad.

Dado que las pruebas clínicas no son perfectas, no es posible utilizarlas correctamente si no se estima la probabilidad previa de daño que presenta el paciente a ser estudiado

¿Que ocurre si nos equivocamos en la apreciación de la probabilidad previa de daño?

- El médico es más capaz de lo que supone para efectuar esa estimación.
- Dicha capacidad es ejercitable y perfectible.
- Un error moderado no tiene implicancias groseras.
- Mayores errores pueden cometerse por no interpretar cabalmente el mensaje de las pruebas.

Evaluación integral de la probabilidad de daño

Una persona tiene ciertas probabilidades de poseer un determinado daño a la salud debido a sus características intínsecas. Por ejemplo: un hombre de 60 años tiene más probabilidades de padecer un cáncer de pulmón si es fumador que si no lo es; si fuma mucho tiene más probabilidades que si fuma muy poco; si fuma

desde hace muchos años tiene mayores probabilidades que si comenzó a fumar hace unos meses. Ahora bien, asistiendo a un hombre, como dijimos de 60 años, que presenta tos, lo interrogamos sobre esos datos y si fuma mucho y desde hace tiempo, le asignaremos una probabilidad moderada de padecer un cáncer de pulmón. Si es el caso contrario, le asignaríamos una probabilidad escasa.

En ambos casos, solicitamos una radiografía de tórax e imaginemos que en ambas placas observamos una mancha oscura. ¿La probabilidad de que se trate de la imagen radiológica de un cáncer de pulmón es similar para ambos casos?. Seguramente que responderá no. Acordará con nosotros en que en el individuo que presenta factores de riesgo, la probabilidad es elevada, mientras que en el otro es baja. Perfecto!!. Usted ha decidido el valor predictivo de una imagen radiológica en función de la probabilidad previa de daño que presentaba el paciente.

Ahora, tratemos de cuantificar esa apreciación.

Si la visualización de una mancha en el campo pulmonar tiene, por ejemplo, un IEP(+) de 7 para cáncer de pulmón y antes de obtener la placa, al paciente de riesgo le habíamos asignado una probabilidad moderada de padecerlo ahora que sabemos que tiene una “mancha”, su probabilidad de daño es mayor. Para calcular la probabilidad actual de daño integraremos la probabilidad previa con la capacidad de la prueba, procedemos de la siguiente manera:

Cálculo de la probabilidad posterior a la aplicación de una prueba:

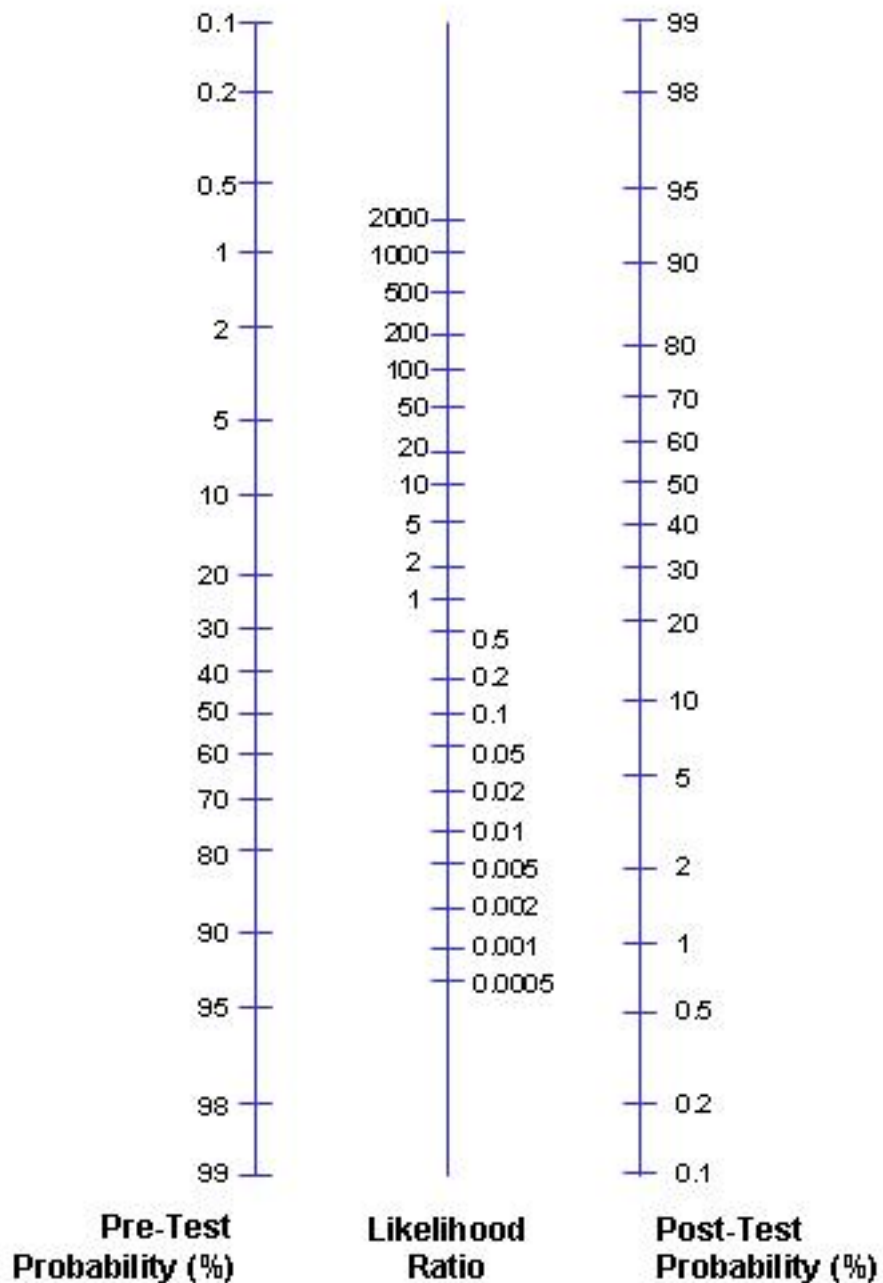
Paso 1: Probabilidad previa / (1 - Probabilidad previa) = Índice pre-prueba

Paso 2: Índice pre-prueba * IEP = Índice pos-prueba

Paso 3: Índice pos-prueba / (1 + Índice pos-prueba) = Probabilidad posterior

Los pasos parecen complejos pues debemos hacer interactuar a una probabilidad con un índice y es matemáticamente incorrecto hacerlo directamente. Por ello, el primer paso es transformar la probabilidad en un índice (índice pre-prueba). Luego hacer interactuar a ambos índices para lograr el índice pos-prueba y por último, transformar el índice resultante en una probabilidad (probabilidad posterior), ya que el médico se maneja con el concepto probabilidad.

Como puede ver para estos cálculos se necesita sólo de una aritmética sencilla, pero sería poco práctico tener que efectuarlos ante cada paciente. Para facilitar la tarea durante el proceso asistencial, hay disponible un nomograma (Fagan TJ. *Nomogram for Bayes theorem. N Engl J Med 1975; 293: 257*) que la torna muy simple y expeditiva.



El nomograma presenta la escala de probabilidades pre-prueba o previas, la del IEP y la de probabilidades pos-prueba o posterior. Trazando una línea que una los dos datos que dispone el médico al recibir el resultado de la prueba (probabilidad

pre-prueba y el IEP de la prueba solicitada), la línea cortará la tercera columna indicando la probabilidad pos-prueba o posterior. Como puede comprobar es sumamente sencillo. Haga la prueba con los casos relatados del presunto cáncer de pulmón.

La probabilidad previa del fumador de padecer cáncer de pulmón, la hemos estimada como moderada a elevada (entre 40 a 60%); la del no fumador, como baja (menos del 20%). Recuerde que dijimos que el IEP de la prueba, es 7. Si procedemos a operar con el nomograma, vemos que la probabilidad posterior para el primer caso es entre 0.8 (82%) a 0.9 (91%), según hayamos considerado una probabilidad previa de 40% ó 60%. Tanto 82% como 91% son muy altas probabilidades de padecer cáncer de pulmón. ¿Notó que en la estimación de la probabilidad previa, tanto 40 ó 60 % no indujeron a diferencias groseras en la evaluación de la probabilidad posterior que alteren la conducta clínica?. La conducta clínica será la misma con 80% ó 90% de probabilidad pos-prueba.

Ahora hagamos el ejercicio para el paciente con probabilidad baja (entre 5 y 19%). Los cálculos nos dicen que la probabilidad pos-prueba se ubica entre 27 y 60%, según se considere una previa de 5 ó 20%, respectivamente. La situación es aquí más compleja, pero dado que se trata de una patología muy grave, aún la menor probabilidad obtenida (27%) es lo suficientemente elevada como para justificar la indicación de otros estudios más complejos o riesgosos y certeros.

Si el paciente hubiera sido una mujer joven, no fumadora, la probabilidad previa de cáncer de pulmón la hubiéramos estimado como muy baja (menos de 5%). La visualización de una mancha, arrojaría una probabilidad posterior de cáncer de pulmón entre el 1 al 22%, considerando una previa entre el 1% y el 4%, respectivamente. La conducta ya no sería la misma que en los casos anteriores, sino que intentaríamos estudiar más las características de la imagen por medio de procedimientos no agresivos.

Como vemos se trataría de una prueba con capacidad suficiente, que nos orienta significativamente para el manejo clínico de los casos, pero no obtenemos respuestas taxativas sobre el estado del paciente, salvo cuando nuestra estimación previa fue elevada. En las demás alternativas, queda un remanente de incertidumbre que el médico deberá manejar con cautela en beneficio del paciente.

Un estudio (Nardone DA y cols. *Usefulness of physical examination in detecting the presence or absence of anemia. Arch Intern Med 1990; 150: 201-4*) calculó el IEP para la estimación de anemia (definida como hematocrito <35% y hemoglobina <11 g/dl en mujeres y <13 g/dl, en hombres), acorde a la palidez facial, la palidez ungueal y la palidez palmar.

Signo clínico	Con anemia (IEP +)	Sin anemia (IEP -)
Palidez facial	3.8	0.6
Palidez ungueal	1.7	0.6
Palidez palmar	2.5	0.5

Visto estos resultados, surge la pregunta, ¿es útil buscar estos signos al realizar el examen clínico?

Muchos de los estudios diagnósticos que solicitamos a nuestros pacientes no informan solamente un resultado normal o anormal, sino que éste puede escalonarse en diversos niveles de anormalidad. Lo que vimos hasta ahora presenta los resultados solamente en dos niveles: normal y anormal. Entonces, ¿cómo apreciar la capacidad operativa de los diferentes niveles de resultados de una prueba cuando esta expresa diferentes niveles de anormalidad?.

Para ello, construimos la tabla de contingencia con las filas necesarias para contener las clases de niveles que dictamine la prueba en análisis. y en las columnas mantendremos los “con” y “sin” daño que es habitual.

Veamos un ejemplo, derivado de un estudio realizado a propósito de observar en que medida la utilización de una tirilla reactiva identificaba proteinuria. El patrón oro aplicado fue la determinación de proteinuria en 24 horas por un medio laboratorial. Observe con atención la manera en que se desarrollaron los cálculos.

Prueba Tirilla reactiva	Proteinuria según laboratorio		Total
	Presente (>0.3 g/L)	Ausente	
3 ó 4 +	12	0	12
2 +	5	2	7
1 +	2	12	14
negativa	3	31	34
Total	22	45	67

$$\text{IEP } 3 \text{ ó } 4 +: (12 / 22) / (0 / 45) = 54$$

$$\text{IEP } 2 +: (5 / 22) / (2 / 45) = 5$$

$$\text{IEP } 1 +: (2 / 22) / (12 / 45) = 0.34$$

$$\text{IEP negativo: } (3 / 22) / (31 / 45) = 0.19$$

Comprobamos que los diferentes niveles de resultados poseen una repercusión clínica claramente diferente. Uno (3 ó 4 +) posee fuerte capacidad para incrementar la probabilidad previa de daño. Dos, resultan muy poco relevantes (2 + y 1 +), mientras que el resultado negativo resulta clínicamente útil para reducir sensiblemente la probabilidad previa de daño.

El IEP incrementa la probabilidad previa de daño en la medida que resulta mayor a 1, en tanto, las reduce en la medida que resulta menor a la unidad.

¿Todos los que observen el resultado de una prueba diagnóstica llegan a la misma conclusión?

Obviamente, no. Y esto usted lo vive en su práctica diaria. Recuerde cuántas veces discutió con sus colegas la conclusión respecto a una imagen radiográfica o un electrocardiograma.

Dado que muchas de las pruebas que se utilizan a diario exigen de eso que se ha llamado “arte”, es importante evaluar la concordancia que existe en los observadores e, incluso, para el mismo observador al evaluar los resultados. A lo primero lo llamamos concordancia o acuerdo interobservador y a lo segundo, acuerdo o concordancia intraobservador. Una buena prueba será aquella que exhiba la mayor concordancia tanta intra como interobservador.

Cómo se evalúa?

Fácilmente podemos intuir que calcularemos en que proporción coincidieron cuando dijeron que había daño y cuántas cuando estimaron que no lo había.

El problema de este sencillo razonamiento es que no estima la influencia del acuerdo por azar. Simplemente por casualidad, habrá casos en que los diagnósticos coincidieron. Por lo tanto, debemos contar con un procedimiento que controle la coincidencia por azar e informe cuánta fue la coincidencia real.

Para ello, existen varios indicadores, pero el más empleado es el propuesto por Cohen (*A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46*) que se llama índice kappa (letra del alfabeto griego). Sus valores van desde el 0 (ausencia absoluta de concordancia) a 1 (concordancia absoluta).

Landis y Koch (*The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174*) propusieron, y es ampliamente usada, la siguiente escala de valoración del índice kappa.

Kappa	grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

Cálculo:

$$\frac{\text{coincidencia observada} - \text{coincidencia esperada}}{1 - \text{coincidencia esperada}}$$

- coincidencia observada = $(a+d) / N$
- coincidencia esperada = $[(a+b)*(a+c) + (c+d)*(b+d)] / N * N$

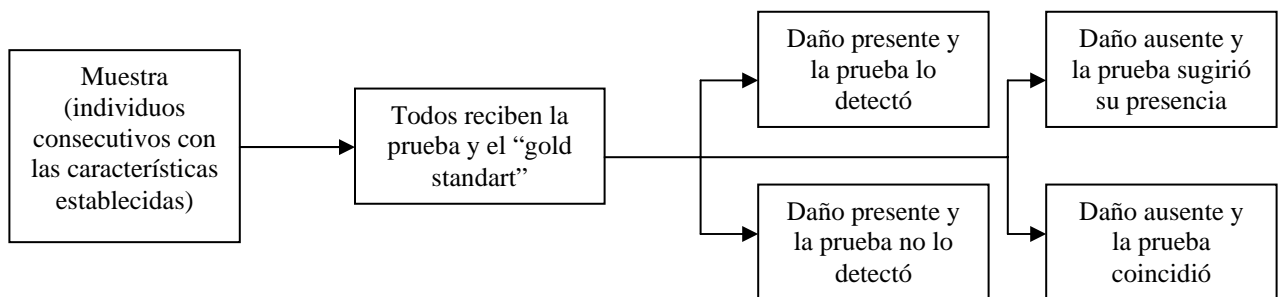
El coeficiente kappa es apropiado para resolver objetivos como el siguiente: evaluar la concordancia de los diagnósticos anatomopatológicos confrontando los producidos por varios patólogos sobre las mismas piezas, cómo así también, los dados por el mismo patólogo sobre las mismas piezas en diferentes ocasiones (variabilidad inter e intra-observador).

El mensaje del coeficiente kappa es más confiable, si muestra que la concordancia es pobre, que si refiere que es elevada. Esto se debe a que si en la muestra estudiada, la proporción de una de las alternativas de calificación del daño es mucho más frecuente que la otra (ej.: gran disparidad entre la proporción de individuos con y sin daño) y para ella la concordancia es adecuada, el resultado

expresará elevada concordancia global, pero en realidad, está informando la concordancia para la alternativa de calificación de mayor frecuencia. Por ello se ha propuesto calcular por separado la concordancia para cada alternativa, permitiendo una visión más equilibrada del comportamiento del método.

Diseño de investigación para evaluar una prueba diagnóstica

El modelo apropiado es un diseño descriptivo, en el cual un grupo de participantes con características determinadas acorde a la prueba a probar son sometidos a la prueba y concomitantemente al “gold standart”.



De esta manera, al cabo del estudio se dispondrá de los datos necesarios para completar la tabla de contingencia, permitiendo calcular todos los indicadores presentados.

Para minimizar los sesgos, se deben considerar algunas premisas claves, a saber:

1. Selección de la muestra: la muestra debe incluir participantes que representen todas las formas clínicas del daño buscado y respetando la proporción en que la misma se presenta en la práctica cotidiana. Aún pruebas de escasa capacidad detectan aceptablemente los casos graves al compararlas con los sin daño. El mayor problema se presenta al intentar detectar a los casos de gravedad moderada a leve. Por ello, la muestra debe obtenerse desde la presentación consecutiva de casos clínicos que reúnan las características que los haría pasibles de la indicación de la prueba. Entre ellos, habrá casos de todas las formas clínicas e incluso casos sin daño.

2. Comparación: la misma debe establecerse con un patrón de referencia apropiado y confiable. El “gold standart” debe reunir esas características y TODOS los participantes deben recibir las dos pruebas (la que se está investigando y el “gold estándar”). Es frecuente observar que si se seleccionó un “gold standart” invasivo y/o costoso, éste se aplique solamente a los que arrojaron un resultado anormal de la prueba en análisis. Este proceder impide que se completen los datos de la tabla de contingencia, ya que no se dispondría de los necesarios para llenar los casilleros c ni d y, por lo tanto, se hace imposible calcular sensibilidad ni

especificidad. A este error de procedimiento (o error sistemático o sesgo) se lo denomina error de verificación (“work-up bias”).

3. Interpretación de los resultados: La lectura del resultado de la prueba analizada debe ser realizado en desconocimiento del resultado del “gold standart”, para evitar el sesgo hacia una coincidencia espúrea.

La guía de lectura crítica

Hemos revisado los indicadores y sus aplicaciones. Tenemos las herramientas básicas. Ahora veamos que le debemos exigir a una investigación sobre la capacidad operativa de una prueba clínica para que sus resultados sean confiables. Las respuestas las hallamos en las guías de lectura crítica. Para la adecuada interpretación de las mismas, usted deberá leer la guía adjunta en la cual hallará la explicación a cada una de esas preguntas. Comprendidas las mismas, lo que le resultará simple dado los conocimientos que adquirió en este módulo, estará capacitado para efectuar el análisis crítico de un artículo sobre la evaluación de una prueba clínica. Adjunto hallará la denominada [Hoja de Trabajo](#), que sintetiza las preguntas de la guía a fin de que le sirva como “check list” y le facilite la tarea de la lectura crítica. En cada uno de los módulos restantes encontrará la correspondiente al tema del mismo.

No piense que por estar publicada una investigación sobre las capacidades de una prueba clínica es confiable. Reid (*Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995; 274: 645*) revisó el período 1978 a 1993, del British Medical Journal, New England Journal of Medicine, The Lancet y del Journal of the American Medical Association buscando artículos que evaluaran pruebas diagnósticas. Rescató 122 y observó que las características de la población incluida estaba bien definida solamente en el 27% de ellos; las de sub-grupos, en el 9%; el control del sesgo de selección, en el 51%; el de información, en el 43%; la precisión de los indicadores, en el 12%; el manejo de los resultados dudosos, en el 26% y la reproducibilidad de los resultados, en el 26%. Como vemos, se justifica plenamente que al leer un artículo de esa naturaleza, usted se encuentre bien capacitado para realizar su análisis crítico.

Bibliografía recomendada

- Sackett DL, Haynes RB. [The architecture of diagnostic research](#). BMJ 2002; 324: 539-41
- Sackett DL, Strauss SE, Richardson WS, Rosenberg W, Haynes RB. Evidence Based Medicine – How to Practice and Teach EBM. Ed. Churchill Livingstone. 2nd. Edition. p. 67-93.
- Knottnerus JA, van Weel C, Muris JWM. [Evaluation of diagnostic procedures](#). BMJ 2002; 324: 477-80.
- Irwig L, Bossuyt P, Glazsiou P, Gatsonis C, Lijmer J. [Designing studies to ensure that estimates of test accuracy are transferable](#). BMJ 2002; 324: 669-71

- Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323:157-62
- Mulherin SA, Miller WC. [Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation](#). Ann Intern Med 2002; 137: 598-602
- Bossuyt P, Reitsma J, Bruns D, Gatsonis C y cols. Towards complete and accurate reporting of studies of diagnostic accuracy: [the STARD initiative](#). Ann Intern Med 2003; 138: 40-4
- Harper R, Reeves B. [Reporting of precision of estimates for diagnostic accuracy: a review](#). BMJ 1999; 318: 1322-3.
- Arkin CF, Wachtell MS. How many patients are necessary to assess test performance?. JAMA 1990; 262:275-8.
- Obuchowski NA. Sample Size Tables For Receiver Operating Characteristic Studies. Am J Roentgenol 2000;175:603-8.
- Steurer J, Fisher JE, Bachmann LM, Koller M, ter Riet G. [Communicating accuracy of test to general practitioners: a controlled study](#). BMJ 2002; 324: 824-6.