Artículo de revisión
Acta Pediatr Mex. 2019 noviembre-diciembre;40(6):342-57.

Estudios de exactitud diagnóstica

Diagnostic accuracy research

Luisa Díaz-García,[1] Isabel Medina-Vera,[1] Silvestre García-de la Puente,[2] Alejandro González-Garay,[1] Chiharu Murata[3]

Resumen

ANTECEDENTES: En los estudios de prueba diagnóstica puede hacerse la distinción en cinco fases que corresponden a diferentes objetivos. La fase III se denomina: estudio de exactitud diagnóstica.

OBJETIVO: Ofrecer a los médicos y a otros profesionales de la salud una introducción actualizada de los conceptos y herramientas estadísticas de estudio de exactitud diagnóstica.

DESCRIPCIÓN: El estudio de exactitud diagnóstica tiene como propósito determinar si la prueba de interés distingue a los sujetos con y sin la enfermedad de entre quienes se sospecha la enfermedad en una situación clínica cotidiana. Para cumplir este objetivo son decisivas: la selección del estándar de referencia válido, evitar los sesgos característicos de esta fase, identificar los parámetros que cuantifican la bondad de la exactitud diagnóstica y efectuar análisis estadísticos adecuados.

CONCLUSIONES: Tanto para el uso de la información como para la realización del estudio de exactitud diagnóstica es necesario distinguir claramente esta fase de estudio como parte de los ensayos diagnósticos. Es importante conocer que la utilidad de la información de las fases II y III es muy diferente. Los usuarios de la prueba diagnóstica deben saber aplicar correctamente los datos locales para que la utilidad diagnóstica generada por la prueba sea adecuada al contexto clínico.

PALABRAS CLAVE: Prueba diagnóstica; exactitud diagnóstica; sesgo; sensibilidad; especificidad; razón de verosimilitud; probabilidad preprueba y posprueba; curva ROC.

Abstract

BACKGROUND: Diagnostic test studies can be distinguished in five stages, corresponding to different objectives. We focus on Phase III, known as diagnostic accuracy research.

OBJECTIVE: To provide physicians and other health professionals with an up-to-date introduction to basic concepts and statistical tools in diagnostic accuracy studies.

DESCRIPTION: A diagnostic accuracy study is carried out to determine if the test of interest distinguishes in a daily clinical situation the subjects who have and do not have the disease among the patients with suspicion of the disease. To achieve this objective, selecting a valid reference standard, avoiding characteristic bias in this phase, identifying parameters that quantify the accuracy of the diagnosis and performing appropriate statistical analyzes are of crucial importance.

CONCLUSIONS: This research phase must be clearly differentiated within the diagnostic study both for the use of information and for diagnostic accuracy studies. In particular, it is important to notice that the usefulness of the information in Phase II and Phase III is very different. Users of diagnostic tests should know how to correctly apply local data so that the results generated by the index test are appropriate for their clinical context.

KEY WORDS: Diagnostic test; Diagnostic accuracy; Bias; sensitivity; Specificity; Likelihood ratio; Pre-test and post-test probability; ROC curve.

[1] Doctor en Ciencias, metodología de la investigación.
[2] Doctor en ciencias, jefe del Departamento, metodología de la investigación.
[3] Maestro en Rehabilitación Neurológica, metodología de la investigación.
Instituto Nacional de Pediatría, Ciudad de México.

Recibido: 14 de diciembre 2018

Aceptado: 14 de octubre 2019

Correspondencia
Chiharu Murata
chiharumurata@gmail.com

Este artículo debe citarse como
Díaz-García L, Medina-Vera I, García-de la Puente S, González-Garay A, Murata C. Estudios de exactitud diagnóstica. Acta Pediatr Mex. 2019;40(6):342-57.
DOI: http://dx.doi.org/10.18233/APM-40No6pp342-3571933

ANTECEDENTES

Los estudios de pruebas diagnósticas abarcan varias etapas, con diferentes tipos de objetivos. David Sackett, R. Brian Haynes y otros precursores de la construcción del paradigma de medicina basada en evidencia establecieron cinco fases de estudio cuyo objetivo corresponde a la pregunta clínica en torno a la elaboración y aplicación de la prueba diagnóstica: comparar la distribución del resultado de la prueba de interés entre los sujetos con y sin la enfermedad (fase I); comparar la morbilidad entre los grupos de sujetos clasificados positivos y negativos de acuerdo con el resultado de la prueba de interés (fase II); estimar los parámetros de utilidad diagnóstica de la prueba de interés en los sujetos en quienes puede sospecharse la enfermedad (fase III); determinar si los pacientes con la prueba de interés tienen mejor pronóstico que quienes padecen la misma enfermedad pero sin hacérseles esta prueba (fase IV); determinar si la implementación de la prueba diagnóstica de interés es benéfica para la salud de los pacientes a un costo adecuado (fase V).[1,2]

Entre estas fases se tiene particular interés en discutir la fase III. Es una parte decisiva de la totalidad del proceso de estudios de prueba diagnóstica, en el sentido de que convierte los datos científicos de la prueba diagnóstica en susceptibles de ser aplicados a una situación clínica real, que merece evaluar su beneficio para el paciente y la sociedad. Es frecuente que se originen problemas debido a la confusión de un estudio fase II y III. Un estudio fase II, expuesto como si fuera el resultado correspondiente a un estudio fase III, suele decepcionar a la comunidad médica. Un estudio fase II puede ser criticado irracionalmente con base en los criterios metodológicos que corresponden a la fase III.[1,2] Por todo esto, para evitar este tipo de confusión, es importante establecer la diferencia conceptual entre las fases II y III.

Este artículo tiene como propósito ofrecer a los médicos y a otros profesionales de la salud una introducción actualizada de los conceptos y herramientas estadísticas de estudio diagnóstico con enfoque en la fase III de las cinco referidas. Los estudios que corresponden a esta etapa son de exactitud diagnóstica (diagnostic accuracy studies).

DESCRIPCIÓN

Características de los estudios de exactitud diagnóstica

Se tratarán las características de esta fase y se comparará con las dos primeras porque es ahí donde surgen ciertas confusiones.

Sackett DL y Haynes RB, así como Haynes RB y You JJ mencionan que la primera y segunda fases del estudio de diagnósticos se llevan a cabo para contestar las siguientes preguntas: *si los pacientes que sufren una enfermedad tienen resultados diferentes en comparación con los individuos sanos con la misma prueba* (fase I); y *si los pacientes con ciertos resultados de la prueba tienen mayor probabilidad de estar padeciendo la enfermedad de interés* (fase II), mientras que la pregunta a la que contestan los estudios de fase III es *si la prueba distingue a los sujetos que tienen y no tienen la enfermedad de interés entre los sujetos en quienes se tiene una sospecha clínica de la enfermedad*.[1,2]

Supóngase que se ha identificado una enzima como potencial biomarcador de una enfermedad y que la investigación se encuentra en la etapa de probar su utilidad en seres humanos.

El primer paso en el proceso de elaboración de una prueba diagnóstica es comparar los valores (por ejemplo, media y desviación estándar) entre los grupos de sujetos con y sin la enfermedad de interés. Si se observa con claridad una di-

ferencia entre los dos grupos puede suponerse su potencial utilidad como biomarcador. Una vez cumplido exitosamente el objetivo de esta primera fase puede invertirse la dirección de la búsqueda y contestar la pregunta de la segunda fase. Para responder esta pregunta, si la medición de la enzima arroja la cuantificación de valores continuos, es conveniente establecer un punto de corte que dicotomice el resultado de la prueba, como se explicará más adelante, por medio de la curva ROC. Gracias a esa dicotomización es posible comparar la proporción de los enfermos entre los grupos de positivo y negativo. Si en el grupo de positivos se encuentra mucho mayor proporción de enfermos y mucho menor proporción de sanos y, al mismo tiempo, mucho menor proporción de enfermos y mucho mayor proporción de sanos en el grupo de negativos, la pregunta de la segunda fase se respondió exitosamente.

Entonces, ¿cuál sería la diferencia entre la fase II y la III? La diferencia reside en la aplicabilidad de los resultados del estudio de la fase III para la práctica clínica cotidiana de la prueba diagnóstica. En la primera y segunda fase los sujetos incluidos en el estudio ya tienen el diagnóstico establecido de enfermos y sanos. Entre estos dos grupos, claramente enfermos y sanos, se comparan la distribución del biomarcador (fase I) y, posteriormente, la proporción de enfermos y sanos entre los positivos y negativos de la prueba (fase II). Así, un biomarcador tiene la diferencia clara entre los grupos (fase I) y los sujetos con resultados positivo y negativo de la prueba y se espera encontrar la diferencia clara de la proporción de enfermos y sanos (fase II). En cambio, en la fase III, se reta la prueba diagnóstica a un conjunto de sujetos con sospecha de la enfermedad de interés con el espectro continuo desde la probabilidad más elevada hasta la menor, atravesando los casos dudosos; por eso esta fase requiere un tamaño de muestra mayor. En esta tercera fase, habitualmente, la capacidad de discriminación observada en la fase II disminuye, por eso los resultados de la segunda fase no son aplicables para diagnosticar a los pacientes en una situación clínica cotidiana. Sin embargo, es razonable efectuar el estudio de la segunda fase porque, con costo y tiempo reducido, puede decidirse si vale la pena llevar a cabo la fase III con el elegido de la prueba diagnóstica que se está investigando.[1-3]

Diseño metodológico de estudios de exactitud diagnóstica

En virtud de las diferencias de cada pregunta que contesta cada fase debe considerarse que el diseño metodológico y los criterios de selección de sujetos de estudio serán distintos para cada una de ellas.

Para la fase I, el diseño será observacional, comparativo y transversal, ya que: estar enfermo o sano no es un factor que se asigna aleatoriamente por parte del investigador (*observacional*); el objetivo del estudio es determinar la diferencia del nivel de la medición que constituye la base del procedimiento de diagnóstico entre los dos grupos, enfermos vs. sanos, (*comparativo*); la medición de la prueba nueva y del estándar de referencia se realizan en una sola ocasión (*transversal*).

La relación que se busca establecer es la asociación de los grupos de individuos (enfermos vs. sanos) con la variable que se pretende utilizar para la prueba diagnóstica, siendo el objetivo de la fase I la comparación de los parámetros estadísticos de la variable que se utilizará para la prueba entre los sujetos que tienen y no tienen la enfermedad. Con respecto a la fuente de datos a analizar puede ser primaria (datos que generan los investigadores es en el estudio planeado) o secundaria (datos ya existentes), es decir, el estudio puede realizarse con el diseño prospectivo o retrospectivo.[2,4,5]

El estudio de la fase II también se realiza como un estudio *observacional*, *comparativo* y *transversal*. Incluso esta fase puede realizarse con el uso de la misma base de datos de la fase I.[1,2] En lo que difiere es la dirección para relacionar las variables: se establecen grupos de comparación en los sujetos, formando los grupos de positivos y negativos, de acuerdo con el nivel de medición de la prueba y se compara la proporción de enfermos y sanos entre los dos grupos, positivos y negativos.

En la fase III, el estudio se realiza, en principio, bajo el diseño *prospectivo*, debido a la necesidad de realizar la medición de variables que generalmente no se espera encontrar en los registros existentes y, además, por el procedimiento metodológico que se requiere para la inclusión de los sujetos en el estudio. Como veremos en la siguiente sección, se requiere incluir a los sujetos en forma consecutiva y que satisfacen los criterios de selección.[2,5]

Ensamble y conducción del estudio

En los estudios de prueba diagnóstica, a través de las primeras tres fases, la elección del estándar de referencia es la parte angular. Cuando se llega a la fase III, la prueba diagnóstica pasa a la situación clínica cotidiana. El estudio de esta fase solo debe realizarse con las pruebas que tuvieron buen rendimiento en las fases previas; el ensamble y conducción es mucho más costoso: se requiere mayor tamaño de muestra con la selección adecuada de los participantes.[1,2]

Selección del estándar de referencia

El estándar de referencia tiene la utilidad de diferenciar entre quienes están enfermos y quienes no lo están; por lo tanto, este estándar es la "realidad" que se utilizará para contrastar la nueva prueba diagnóstica. Para poder seleccionar correctamente este estándar es necesario definir con claridad la enfermedad y sus criterios diagnósticos y de referencia a estudiar. Por lo general, el estándar de referencia es una prueba, pero si no se dispone de una prueba de referencia puede ser un constructo o, bien, puede ser la evolución del paciente a través del tiempo.[1-5]

Selección de la prueba

La prueba diagnóstica a estudiar debe evaluarse para determinar su exactitud (*diagnostic accuracy*); es decir, la capacidad de la prueba para distinguir entre quienes tienen la enfermedad y quienes no; para esto se estiman los parámetros que caracterizan la exactitud diagnóstica. La aceptabilidad de la validez de la nueva prueba dependerá de la enfermedad estudiada y de las condiciones reales en el medio en que será aplicada. La prueba no debe ser una parte constituyente del estándar de referencia porque su utilidad podría sobrevalorarse.[1-5]

Selección de participantes

Los participantes en quienes se evalúa una nueva prueba no deben diferir, sustancialmente, de la población a la que se aplicará en la práctica clínica, incluido un espectro extenso de la enfermedad; es decir, con participantes en diferentes etapas de evolución y gradiente de la enfermedad, incluso participantes sanos; es importante contar con una cantidad importante de participantes en cada estrato. Además, en este espectro de participantes deben incluirse los individuos con diagnósticos difíciles de establecer la diferencia (diagnóstico diferencial) con otros pues ello permite determinar los falsos positivos.[3-6]

Conducción del estudio

Claramente debe definirse cuándo se considerarán la prueba diagnóstica y el estándar de referencia positivos o negativos. Para calcular los estimadores de los parámetros de utilidad diagnóstica se requiere la comparación de dos

variables nominales dicotómicas para la construcción del cuadro de 2 x 2 (tabla de contingencia).

A todos los participantes debe aplicárseles el estándar de referencia y la nueva prueba diagnóstica de manera simultánea o con un intervalo pequeño entre ellas. La finalidad de esto es que la evolución de la enfermedad no cambie durante el tiempo de espera entre la realización de las pruebas. Además, es decisivo clasificar de manera independiente los casos por el estándar de referencia y por la prueba diagnóstica. El cegamiento de los evaluadores es un requerimiento metodológico fundamental para ambos criterios de clasificación.[1,2,4,5]

Sesgos

En todos los estudios epidemiológicos deben considerarse los posibles sesgos inherentes al diseño que se utilizará para poder minimizarlos. En la fase III de los estudios de prueba diagnóstica, los sesgos más relevantes son:

El ***sesgo de espectro*** puede aparecer cuando la prueba de interés se lleva a cabo en una población de estudio con un espectro clínico diferente al que será ocupado en la práctica cotidiana, ejemplo que al momento de calibrar la prueba se incluyan sujetos con estadios más avanzados de la enfermedad. Esto suele ocurrir cuando las pruebas se ejecutan en centros hospitalarios. Es necesario poner especial cuidado en el espectro clínico que se requiere cubrir con la prueba para que éste quede cubierto en el momento de hacer el estudio.[1, 2, 4-6]

Los ***sesgos de selección*** se dan cuando existe una relación entre el resultado de la prueba y la probabilidad de ser incluido en la población de estudio. Si la cuantificación del antígeno prostático en un paciente es alta es mucho más probable que sea "preseleccionado" para biopsia. Si como resultado de esa preselección "se define" la prueba; es decir, con sujetos preseleccionados la prueba tendrá mayor sensibilidad pero menor especificidad. Es indispensable tener especial cuidado en el mecanismo de selección de los sujetos de estudio; la muestra de participantes debe ser representativa de la población en la que se aplicará la prueba. Este tipo de sesgo puede ocasionar una sobreestimación de la sensibilidad o de la especificidad.

Uno de los sesgos de selección más común en este tipo de estudio es el ***sesgo de confirmación diagnóstica*** y sucede cuando el estudio se limita a los participantes a los que ya se les realizó la prueba de referencia; esto porque quizá estos participantes sí tengan la enfermedad y, por tanto, los resultados positivos se ven sobrerrepresentados (sobreestimación de la sensibilidad) y los resultados negativos se subestiman (subestimación de la especificidad). Con frecuencia, los sesgos de espectro y de selección están relacionados y pueden afectar todas las medidas de discriminación.

El sesgo de interpretación de las pruebas o sesgo de sospecha diagnostica. Los observadores que aplican la prueba diagnóstica y la prueba de referencia deben emitir los resultados de forma independiente y cegada, para evitar que los resultados de una prueba puedan influir en la interpretación de la otra. Debe considerarsae cuando los resultados de las pruebas dependen de la destreza del observador y si las pruebas se realizan por uno o más observadores porque puede haber variabilidad interobservador; por esto se recomienda que antes del inicio del estudio se capacite adecuadamente a los clínicos que aplicarán o interpretarán las pruebas. [2,4,5]

Sesgo de mala clasificación. El sesgo de mala clasificación en los estudios de prueba diagnóstica es el que ocurre cuando las observaciones clínicas o las técnicas de laboratorio son imperfectas. Por lo tanto, algunos participantes sanos

se clasificarán erróneamente como enfermos, mientras que otros que padecen la enfermedad no serán identificados.

La distorsión de la estimación de la sensibilidad y especificidad de la prueba, que son distintas para cada grupo, sesga los resultados en cualquier dirección y la existencia o ausencia de una verdadera asociación puede quedar enmascarada, disminuida o aumentada o, quizá, se encuentre un efecto que realmente no existe.[2,4,5]

Métodos estadísticos

Se exponen los parámetros básicos para cuantificar la exactitud diagnóstica y las técnicas estadísticas utilizadas para su estimación. En el apéndice se aportan ejemplos del análisis estadístico, utilizando los datos de Alistair F. Smith, "*Diagnostic value of serum-creatine-kinase in a Coronary-Care Unit*", un artículo clásico de estudio de diagnóstico de infarto de miocardio publicado en *Lancet* .[7]

Parámetros de prueba diagnóstica

Los análisis estadísticos en estudios de exactitud diagnóstica están destinados a evaluar qué tan correctamente clasifica un procedimiento diagnóstico a pacientes con y sin una enfermedad. Para esta evaluación se dispone de varios parámetros que cuantifican el rendimiento de la prueba diagnóstica.

El primer paso para calcular estos parámetros es construir una "tabla de contingencia" de dos columnas y dos renglones; su combinación genera cuatro casillas. Las columnas corresponden a los sujetos "enfermos" y "sanos" de acuerdo con el estándar de referencia y los renglones corresponden a los resultados "positivo" y "negativo" arrojados por la prueba diagnóstica en evaluación. En la **Figura 1** se aprecia cómo convencionalmente se asignan a las 4 casillas las

A

		Diagnóstico de referencia +	Diagnóstico de referencia -	
Prueba	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	a+b+c+d

B

		Diagnóstico de referencia +	Diagnóstico de referencia -	
Prueba	+	Verdaderos positivos(VP)	Falsos positivos(FP)	Total casos con prueba positiva
	-	Falsos negativos(FP)	Verdaderos negativos(VN)	Total casos con prueba negativa
		Total enfermos	Total sanos	Gran total

Figura 1. Tabla de contingencia 2 × 2. A) Nomenclatura de las casillas de la tabla y B) lo que implica cada casilla de la tabla.

letras *a, b, c* y *d*: a) **Verdaderos positivos (VP)**, porque son sujetos enfermos y la aplicación de la prueba resulta positiva; b) **Falsos positivos (FP)**, porque son sujetos sanos pero la prueba resulta positiva; c) **Falsos negativos (FN)**, porque son sujetos enfermos, pero la prueba resulta negativa; y d) **Verdaderos negativos (VN)** porque son sujetos sanos y la prueba fue negativa.[2]

Al sumar la frecuencia de las dos columnas y los dos renglones se obtienen las "frecuencias marginales". Las dos frecuencias marginales de los renglones ("*a+b*" y "*c+d*") son totales de los casos que resultaron "positivos" y "negativos"; las dos frecuencias marginales de las columnas ("*a+c*" y "*b+d*") son totales de "enfermos" y "sanos" definidos por el estándar de referencia. La suma de las frecuencias marginales de renglones y de columnas será el "gran total". La proporción de la frecuencia marginal de la columna de

"enfermos" es la prevalencia de la enfermedad estudiada y la proporción de la frecuencia marginal de la columna de "sanos" es el complemento de la prevalencia (1-prevalencia). Hay que tomar en cuenta que para que esta prevalencia sea un estimador adecuado es necesario obtener los datos por conceptualización y procedimientos adecuados de muestreo.[6]

Los parámetros de prueba diagnóstica más básicos son: sensibilidad, especificidad, complemento de la sensibilidad, complemento de la especificidad, valor predictivo positivo y valor predictivo negativo. Todos estos parámetros son proporciones (cuantificación de p /[p + q] donde el numerador es parte del denominador) en una subpoblación de la totalidad: subpoblación de "enfermos", "sanos", "sujetos con el resultado positivo de la prueba" o "sujetos con el resultado negativo de la prueba". Al tratar estas proporciones en una subpoblación como una probabilidad, se dice que son probabilidades condicionadas porque se calcula la probabilidad, bajo una condición que tiene una probabilidad para suceder.

Sensibilidad. Probabilidad de que la prueba sea positiva puesto que se tiene la enfermedad; se refiere al número de casos que caen en la casilla (verdaderos positivos), dividido entre el total de los enfermos: a/(a + c).

Especificidad. Probabilidad de que la prueba sea negativa puesto que no se tiene la enfermedad. De manera similar a la sensibilidad es la proporción de verdaderos negativos dividida entre el total de los sanos: d/(b+d).

Proporción de falsos negativos. Probabilidad de que la prueba sea negativa porque se tiene la enfermedad. Es igual a falsos negativos entre el total de los enfermos: c/(a + c). Es frecuente que se llame "tasa de falsos negativos (false negative rate)"; sin embargo, de acuerdo con la definición del término, no es una "tasa" sino una "proporción".[8] Esta proporción se calcula verticalmente en la tabla de contingencia presentada, porque es una probabilidad condicionada determinada dentro de la subpoblación "enfermos". Como esta proporción es la parte complementaria de la sensibilidad también suele expresarse como "1 – Sensibilidad".

Proporción de falsos positivos. Probabilidad de que la prueba sea positiva en virtud de que no se tiene la enfermedad. Es igual a falsos positivos entre el total de los sanos: b/ (b + d). Igual que la proporción de falsos negativos, es frecuente que se denomine "tasa" aunque en realidad es una "proporción". Esta proporción se calcula verticalmente en la tabla de contingencia presentada. Como esta proporción es la parte complementaria de la especificidad, suele expresarse también como "1– Especificidad".

La sensibilidad y especificidad son parámetros que expresan el rendimiento de la prueba diagnóstica y ofrecen las bases para calcular otros parámetros importantes.[8,9] Carece de sentido interpretar la sensibilidad sin tomar en cuenta la especificidad e interpretar la especificidad sin considerar la sensibilidad. La sensibilidad se eleva a lo máximo (100%) si se sacrifica la especificidad y lo mismo puede hacerse con la especificidad. Puesto que la sensibilidad es la proporción de sujetos positivos en la subpoblación de enfermos, si se clasifica a todos los sujetos objeto de la prueba como "positivos", la sensibilidad será de 100%. Pero, lógicamente, bajo esta condición, la especificidad será 0%. Con la especificidad ocurre exactamente lo mismo, si se clasifica a todos los sujetos evaluados por la prueba como "negativos".

Con una prueba típicamente de tamizaje si el procedimiento cuenta muy alta sensibilidad, con adecuada especificidad, en los sujetos con el resultado negativo podrá descartarse la en-

fermedad y la probabilidad de cometer error es baja porque en esta prueba la probabilidad de falso negativo es baja.

De manera simétrica, en una prueba de diagnóstico confirmatorio, si el resultado es positivo, con elevada probabilidad el sujeto está enfermo. Estas reglas de interpretación de los resultados se denominan "SnNout" y "SpPin", acrónimo inglés de ***S**ensitive test **N**egative ⟶ to rule **out** the diagnosis* y ***Sp**ecific test that is **P**ositive serves to rule **in** the diagnosis*.[2]

A pesar de que la sensibilidad y especificidad, y sus complementos, la proporción de falsos negativos y positivos son parámetros fundamentales para cuantificar la propiedad de una prueba; por sí solos no aportan la información que el clínico realmente necesita porque lo que requiere saber el médico que aplica la prueba diagnóstica no es el rendimiento de ese instrumento diagnóstico, sino qué tan probable es que el sujeto esté enfermo (o no) en virtud del resultado positivo (o negativo) con esa prueba. Los parámetros que estiman estas probabilidades son valores predictivo positivos y negativos.[2,5,10]

Valor predictivo positivo (VPP). Probabilidad de que un paciente tenga la enfermedad si la prueba resulta positiva. Es igual a los verdaderos positivos entre el total de los casos con prueba positiva: a/(c + d).

Valor predictivo negativo (VPN). Probabilidad de que un paciente no tenga la enfermedad si una prueba resulta negativa. Es igual a los verdaderos negativos entre el total de los casos con prueba negativa: d/(c + d).

El problema que hay que considerar con los valores predictivos es que cambian de acuerdo con la prevalencia; el *VPP* aumenta o disminuye en forma proporcional al aumento o disminución de la prevalencia. En cambio, el *VPN* aumenta o disminuye en sentido inverso a los cambios en la prevalencia. Entonces, aunque un artículo de una prueba diagnóstica reporte el *VPP* o *VPN* muy elevados, no necesariamente esos resultados serán aplicables a los pacientes del médico que leyó ese artículo, porque la prevalencia de la enfermedad de interés en los pacientes atendidos en su hospital puede ser muy diferente de la población de los pacientes estudiados en el artículo que generó los estimadores de *VPP* y *VPN*.[2, 5]

Otros parámetros que tienen una participación decisiva en el estudio de exactitud diagnóstica: la razón de verosimilitud positiva y negativa. La diferencia de estos parámetros frente a los explicados anteriormente es que son comparaciones de dos probabilidades. El término "razón" refiere una comparación de dos cantidades en forma de cociente y el numerador y denominador no constituyen la relación de parte y total.

***Razón de verosimilitud positiva* (*RV+*)**. Razón de "sensibilidad/(1-especificidad)". Como "1-especificidad" es la proporción de falsos positivos es una cuantificación que contesta la pregunta "¿cuántas veces mayor es la probabilidad de verdadera positividad en comparación con la falsa positividad?". El valor mayor que 1 de este cociente significa que la probabilidad de la verdadera positividad es mayor que la falsa positividad, al aplicar la prueba diagnóstica. Si RV+ es igual que 1, eso quiere decir que con la misma probabilidad genera la verdadera positividad y falsa positividad, entonces, la prueba no tiene ninguna utilidad. Convencionalmente se considera que es útil al ser LR+ 2 y muy útil si LR+ ≥ 10.[9]

***Razón de verosimilitud negativa* (*RV-*)**. Razón de "(1-sensibilidad)/ especificidad". Como "1-sensibilidad" es la proporción de falsos positivos es una cuantificación que contesta a la pregunta "¿cuántas veces menor es la probabilidad de falsa negatividad en comparación con la verdadera negatividad?". El valor menor que 1 de

este cociente significa que la probabilidad de la falsa negatividad es menor que la verdadera negatividad al aplicar la prueba diagnóstica. Si RV- es igual que 1, eso quiere decir que con la misma probabilidad genera la falsa negatividad y verdadera negatividad; entonces, la prueba no tiene ninguna utilidad. Convencionalmente se considera que es útil al ser LR- < 0.5 y muy útil si LR- < 0.1.[9]

RV+ y RV- por sí solas son buenas cuantificaciones para evaluar la exactitud diagnóstica de una prueba; sin embargo, su verdadera potencia está en su aplicación para calcular la probabilidad de que un paciente tenga la enfermedad de acuerdo con el resultado de la prueba diagnóstica, y determinar la diferencia que genera en las probabilidades antes y después de conocer el resultado de la prueba, que precisamente es cuantificar la utilidad de una prueba diagnóstica.

Probabilidad preprueba y probabilidad posprueba

Se refiere con el término "probabilidad preprueba" al grado que se sospecha tener la enfermedad que establece un clínico con respecto a su paciente sin aplicar la prueba diagnóstica. Para establecer este grado de sospecha, un médico utilizará todos sus conocimientos y experiencia clínica: datos epidemiológicos, clínicos y de laboratorio e intuición como experto. [2,4,9-11]

Una vez establecida esta probabilidad, una síntesis de datos objetivos y juicio subjetivo, puede ponderarse por la RV+ o RV-, dependiendo de cómo resultó la aplicación de la prueba se puede calcular la probabilidad posprueba: la probabilidad de tener la enfermedad. Si la aplicación de una prueba diagnóstica genera una diferencia importante entre la probabilidad preprueba y posprueba de tener o no tener la enfermedad puede decirse que la utilidad de esa prueba diagnóstica es grande. El procedimiento para generar la probabilidad posprueba es el siguiente:

Momio de probabilidad preprueba × RV = momio de probabilidad posprueba

Momio es una cuantificación que se emplea con frecuencia en el manejo de probabilidad, que se define como una razón; es decir, una comparación de dos valores por su cociente, de la probabilidad de ocurrir un evento sobre su complemento. Esta relación se expresa algebraicamente como:

$$\text{Momio} = \frac{\text{Probabilidad de estar enfermo}}{1 - \text{Probabilidad de estar enfermo}}$$

Al despejar "*Probabilidad de estar enfermo*" esta ecuación se obtiene la siguiente expresión:

$$\text{Probabilidad de estar enfermo} = \frac{\text{Momio}}{1 + \text{Momio}}$$

Al utilizar estas dos ecuaciones puede hacerse la conversión de probabilidad ⟶ momio y momio ⟶ probabilidad. Así que, a partir de la determinación de la probabilidad preprueba, por la mediación de la aplicación de una prueba diagnóstica, se obtiene la probabilidad posprueba del paciente en el siguiente procedimiento:[2,4,5,10]

1. Una vez determinada la probabilidad preprueba de un paciente por el médico tratante
2. esta probabilidad se convierte en el momio preprueba;
3. se somete el paciente a una prueba diagnóstica con el RV conocida y se obtiene el resultado,
4. se multiplica el momio preprueba por RV+ o RV- según el resultado de positividad y negatividad;
5. se obtiene el momio posprueba; y

6. se convierte este momio posprueba en la probabilidad.

Mediante este procedimiento puede evaluarse la utilidad de la prueba diagnóstica en: estudio por medio del cambio que genera entre la probabilidad preprueba y posprueba. El ejemplo del cálculo algebraico se presenta en la sección de apéndice.

Intervalo de confianza

En la sección anterior se expusieron los parámetros básicos de estudios de exactitud diagnóstica. Ahora se llamará la atención con el término "parámetro" porque es un término que está definido técnicamente. La International Epidemiological Association define el vocablo "parameter" como: *In statistics and epidemiology, a measurable characteristic of a population that is often estimated by a statistic, e.g., mean, standard deviation, regression coefficients*.[12] Es decir, es una caracterización de una propiedad en la población. Un estudio se efectúa para conocer la característica de una población; sin embargo, salvo estudios de censo en los que se obtiene la medición de todos los elementos que constituyen la población, nunca podemos conocer directamente esos "parámetros". Lo que podemos hacer es estimar los parámetros de nuestro interés por medio del uso de una (o más) muestra (s). Aquí se utiliza el término "muestra" como un término técnico refiriéndose a "una parte de la población".

Al llevar a cabo un estudio, el investigador procura generar un buen estimador del parámetro con base en una muestra representativa, sin embargo, el estimador del parámetro de interés siempre sera susceptible de la aleatoriedad y puede disceptar del valor del parámetro. Si llamamos este estimador que se genera apuntando solo un valor como "*estimador puntual*" podemos reconocer otro tipo de estimador que es el "*estimador por intervalo*", el cual es llamado "*intervalo de confianza*".

Un intervalo de confianza es un rango acotado por el límite inferior y superior en el que se espera encontrar el verdadero valor del parámetro, es decir, el valor poblacional de la medición de interés con cierto nivel de confiabilidad. Los límites inferiores y superiores se determinan sumando y restando al estimador puntual la magnitud del margen de error que se obtiene por la multiplicación de una cantidad denominada "*error estándar*". De acuerdo con el valor utilizado para multiplicar el "*error estándar*" puede cambiarse el nivel de confianza. Al multiplicar el error estándar por 1.96 veces se obtiene el margen de error (o precisión de la estimación) que corresponde al nivel de confianza de 95%.[10]

El intervalo de confianza de 95% de cada uno de los parámetros que presentamos en la sección anterior se puede calcular sumando y restando ***1.96 × el error estándar*** al estimador puntual del parámetro de interés. La expresión algebraica para calcular el error estándar, intervalo de confianza de 95%, así como la manera de calcular el tamaño de la muestra con y sin información de la prevalencia de la enfermedad de interés se señala en el **Cuadro 1.** La expresión $\hat{P}(1-\hat{P})$ es el producto de la proporción (probabilidad) que corresponde a una categoría y $1-\hat{P}$ es su complemento. Con este producto se obtiene la varianza de una variable categórica dicotomizada. Como la sensibilidad y especificidad son proporciones, la letra $\hat{P}$ se sustituye por un valor hipotético de la sensibilidad o la especificidad.

Si hay una prueba diagnóstica con la que esperamos una sensibilidad, digamos, de 85%, el producto $0.85 \times 0.15 = 0.1275$ es la varianza estimada. Dividiendo este valor por un tamaño de muestra y calculando su raíz cuadrada se obtiene el error estándar. Si contamos con 49 sujetos enfermos:

Cuadro 1. Expresión algebraica de error estándar (EE), intervalo de confianza de 95% (IC95%), tamaño de muestra del grupo de enfermos para determinada precisión del IC95% con y sin la información de la prevalencia de la enfermedad de interés para el diagnóstico

Error estándar (*EE*)	Intervalo de confianza de 95% (IC95%)	n para IC95%, cuando la prevalencia es conocida	n para IC95%, cuando la prevalencia es desconocida
$EE = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$	$IC95\% = \hat{\theta} \pm d$ $IC95\% = \hat{\theta} \pm 1.96 \times \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$	$n = \frac{1.92^2 \times \hat{P}(1-\hat{P})}{d^2}$	$n = \frac{1.96^2 \times \hat{P}(1-\hat{P})}{d^2 \times \widehat{Prev}}$

$\hat{P}$: estimador puntual de la proporción de interés, $\hat{\theta}$: estimador puntual del parámetro de interés,
d: mitad del IC (margen de error, precisión), n: tamaño de muestra del grupo de enfermos clasificados por el estándar de referencia, $\widehat{Prev}$: estimador de la prevalencia de la enfermedad

$$EE = \frac{0.1275}{49} = 0.051$$

El error estándar es 0.051. Al multiplicar este valor por 1.96 da el valor de 0.10. Vamos a suponer que la misma muestra había generado el estimador puntual de la sensibilidad como 0.82, entonces,

$IC95\% = 0.82 \pm 0.1$

Por lo tanto,

Límite inferior (LI) y límite superior (LS) del IC95% se determinan: 0.82 – 0.10 = 0.72 y 0.82 + 0.10 = 0.92, respectivamente, por lo que con esta prueba la confianza al 95% podemos estimar que el valor verdadero de la sensibilidad se encontrará entre 72 y 92%.

Curva ROC

Hay pruebas diagnósticas que se basan en las variables numéricas: concentraciones de glucosa, puntaje de un instrumento de evaluación de depresión, valor del antígeno específico prostático, etc. En las fases II y III se requiere un procedimiento para convertir las variables numéricas en variables categóricas en relación con el estándar de referencia. En este procedimiento se utiliza una herramienta matemática llamada "Curva ROC". "ROC" es la sigla de "Receiver Operating Characteristic [Característica Operativa del Receptor]". Este nombramiento poco médico tiene su origen en la ingeniería que desarrolló la tecnología del uso del radar para discriminar correctamente la señal que detecta la presencia de submarinos enemigos (verdadero positivo) frente al ruido falso (falso positivo) y ausencia del enemigo (verdadero negativo) frente a la falsa alarma de la presencia del enemigo (falso positivo).[10]

Si para un conjunto de sujetos en quienes se sospecha una enfermedad, si se aplican simultáneamente un examen médico que arroja resultados numéricos y además el procedimiento del diagnóstico que sirva como estándar de referencia, es posible calcular una serie de valores de sensibilidad y especificidad; cada valor numérico de los pacientes es un "punto de corte". De acuerdo con la diferencia de cada "punto de corte" se generarán las combinaciones de sensibilidad y especificidad. Éstas son las cuantificaciones de las características antagónicas. Si se define el punto que optimiza esta relación de sensibilidad-especificidad como la combina-

ción que maximiza la suma de "sensibilidad + (1-especificidad)", es posible proponer el punto que optimiza estos dos valores. Esta suma se llama "Índice de J de Youden" y se emplea para "operacionalizar" la determinación del punto de corte óptimo.[2,6] En la **Figura 2** se aprecia un ejemplo de la aplicación de la curva ROC.

		Infarto del miocardio		
		+	-	
Prueba	+	215 (a)	16 (b)	231 (a+b)
	-	15 (c)	114 (d)	129 (c+d)
		230 (a+c)	130 (b+d)	360 (a+b+c+d)

Figura 2. Los datos de la utilidad del valor de CK para el diagnóstico de infarto del miocardio (Smith AF, 1967) reorganizados en una tabla de contingencia 2 × 2. El punto de corte de CK≥*80* fue establecido por el autor de acuerdo con la evaluación visual.

Tamaño de la muestra

La necesidad e importancia de calcular el tamaño de la muestra en la etapa de planeación de estudios de diagnóstico se han señalado desde hace tres décadas.[13-16] No obstante, en diferentes estudios de revisiones recientes de la bibliografía médica se menciona que es poco común que los estudios de exactitud diagnóstica reporten el procedimiento para determinar el tamaño de muestra adecuadamente.[17,18]

Por un lado, hay muchos de esos trabajos en los que se determina la cantidad de sujetos enfermos y sanos a estudiar, sobre todo por la disponibilidad y, por otro lado, hay varios otros en los que se decide el tamaño de muestra solo replicando una cantidad similar de los que se incluyeron en otros estudios de temas similares previamente publicados.[17] Como consecuencia de esta tendencia, sobre todo en estudios de exactitud diagnóstica, los estudios publicados cuentan con un tamaño de muestra insuficiente que no permite que el estudio genere resultados con precisión suficiente.[18] Una razón por la que prevalece esta práctica no deseable con respecto al cálculo del tamaño de muestra es la diversidad de los parámetros, objetivos y diseños del estudio del tema de exactitud diagnóstica.[18]

El cálculo del tamaño de la muestra más sencillo es el que consiste en determinar la cantidad de casos necesarios para construir un intervalo de confianza de cierto nivel, digamos, de 95%. Cuando se introdujo el concepto del IC nos referimos al **Cuadro 1** e hicimos un pequeño ejercicio de calcular el error estándar y el intervalo de confianza de 95%. En la tercera columna del mismo cuadro se encuentra una ecuación con el título de " para IC95%, cuando la prevalencia es conocida". Es, más que nada, la expresión obtenida despejando la *n* de la fórmula del IC95%. Ahora suponemos que queremos determinar el tamaño de muestra necesaria para poder estimar la sensibilidad de la prueba diagnóstica con la que creemos que su sensibilidad sea alrededor de 0.85, construyendo un IC95% con una precisión de ± 0.10. Sustituyendo $\hat{P}$ por el valor hipotético de 0.10 de sensibilidad y, la magnitud de precisión que corresponde a una mitad del rango de IC por 0.10, entonces,

$$n = \frac{1.96^2 \times (1 - 0.85)}{0.10^2} = 48.98 \approx 49$$

Nótese que, en este ejemplo, el tamaño de muestra que se estimó es precisamente la que utilizamos para calcular el IC95%. Es lógico porque la parte que define la precisión, en la expresión de la columna de IC95%, es una transformación algebraica de la siguiente ecuación sobre d.

$$d = 1.96 \times \sqrt{\frac{\hat{P}\,(1 - \hat{P})}{n}}$$

Pero, si el asunto es tan sencillo, no es comprensible la dificultad que mencionamos al iniciar esta sección. Ahora vea la última columna del **Cuadro 1**. Las ecuaciones que se encuentran en las dos últimas columnas son para calcular el tamaño de muestra del grupo de enfermos clasificados por el estándar de referencia; la diferencia es que en la última "*n* para IC95%, cuando la prevalencia es desconocida" tiene un factor de "prevalencia de la enfermedad" en el lado del denominador. Como la prevalencia es una proporción, este valor debe ser mayor que 0 (si es 0, eso quiere decir que no existe ningún enfermo y ese grupo deja de existir) y menor que 1 (por la misma razón, pero, con la prevalencia de uno, ahora dejará de existir el grupo de sanos).

Entonces, si la prevalencia de la enfermedad es pequeña, en el estudio se requerirá el tamaño de muestra grande y, si la prevalencia de la enfermedad es grande, se requerirán menos casos para el grupo de enfermos. La diferencia de estas dos fórmulas corresponde a los diseños típicos de los estudios de exactitud de la fase II y fase III. En la fase II el propósito del estudio es conocer si la "positividad" y "negatividad" de la prueba de interés claramente se asocian con el diagnóstico por el estándar de referencia, "enfermo" y "sano". Un diseño adecuado sería de "casos (enfermos) y controles (sanos)", entonces, la prevalencia de la enfermedad es conocida por el investigador porque la razón de los enfermos comparada con los sanos es la parte del diseño que el investigador debe determinar. En este caso no se aplica el factor "prevalencia" en el denominador del cociente. En la fase III, de manera contrastante, el diseño típico es de cohorte comparativo prospectivo y, con este diseño, los investigadores no tienen conocimiento de la prevalencia de la enfermedad porque esa proporción es algo que se determina al terminar de incluir los últimos sujetos en los dos grupos. En estudios de exactitud, fase III, el diseño común es prospectivo y se recomienda, desde el punto de vista del diseño de muestreo, incluir los sujetos enfermos y sanos conforme llegan los nuevos casos.

La integración del factor prevalencia la sugirió Nancy M. Fenn Buderer en 1996 y actualmente está establecida como un procedimiento necesario para calcular el tamaño de muestra de estudio. La distinción adecuada de estas dos fórmulas exige cierto nivel de conocimiento metodológico y este tipo de cuestiones puede causar dificultad para realizar el cálculo del tamaño de muestra en estudios, sin embargo, problemas más complejos surgen cuando los investigadores se enfrentan a situaciones en las que se conjugan la factibilidad, diseños de muestreo y manejo de modelos estadísticos con ajuste multivariado. Para combatir el sesgo de espectro es necesario elaborar el diseño de muestreo adecuado y para integrar en los modelos estadísticos las covariables y las interacciones entre esas variables debe contarse con el conocimiento de modelos estadísticos. Ante una situación multivariada hay algunos autores que ofrece soluciones con base empírica o por simulaciones.

CONCLUSIÓN

Un estudio de prueba diagnóstica abarca una serie de fases a través de las cuales se evalúa la utilidad de la prueba de interés. En este artículo nos enfocamos en la fase III de este tipo de estudio. En esta fase se genera la información que los clínicos necesitan para aplicar las pruebas diagnósticas para atender a sus pacientes. Para lograr generalizar y dar aplicabilidad en una situación clínica real, el estudio de la fase III debe llevarse a cabo con la definición clara de la población con la que se estiman los parámetros de prueba diagnóstica, procurando minimizar los sesgos. Los usuarios de la prueba diagnóstica deben saber aplicar correctamente los datos locales para que sean adecuados a las utilidades

diagnósticas generadas por la prueba para su contexto clínico.

APÉNDICE: Ejemplos de cálculo de los estimadores de parámetros

Creatina cinasa (CK) es una enzima que participa en el metabolismo energético durante la contracción muscular. Hay fuga de esta enzima en caso de lesionar los músculos y en el miocardio se distribuye con mayor proporción su isoenzima tipo MB, por lo que esta última puede servir como biomarcador de infarto de miocardio. "Diagnostic value of serum-creatine-kinase in a Coronary-Care Unit" de Smith AF[3] es un artículo clásico que reportó su utilidad diagnóstica para identificar a los pacientes que requieren atención en unidades de cuidados coronarios. El artículo reporta sólo los resultados que corresponden a la fase I, pero los datos presentados permiten reconstruir la base de datos, por lo que se generaron los resultados que corresponden a la fase III a partir de este artículo.

De acuerdo con los datos que se resumen en la tabla de contingencias de 2 x 2 (**Figura 2**), los parámetros mencionados pueden calcularse de la siguiente manera:

$$\text{Sensibilidad} = \frac{\text{verdaderos positivos}}{\text{total enfermos}} = \frac{a}{a+c} = \frac{215}{230} = 93.5\%$$

$$\text{Especificidad} = \frac{\text{verdaderos negativos}}{\text{total sanos}} = \frac{d}{b+d} = \frac{114}{130} = 87.7\%$$

$$\text{Proporción de falsos negativos} = \frac{\text{falsos negativos}}{\text{total enfermos}} = \frac{c}{a+c} = \frac{15}{230} = 6.5\%$$

$$\text{Proporción de falsos positivos} = \frac{\text{falsos positivos}}{\text{total sanos}} = \frac{b}{b+d} = \frac{16}{130} = 12.3\%$$

$$\text{VPP} = \frac{\text{verdaderos positivos}}{\text{total casos con prueba positiva}} = \frac{a}{a+b} = \frac{215}{231} = 93.1\%$$

$$\text{VPN} = \frac{\text{verdaderos negativos}}{\text{total casos con prueba negativa}} = \frac{d}{c+d} = \frac{114}{129} = 88.4\%$$

$$\text{RV} + = \frac{\text{sensibilidad}}{1 - \text{especificidad}} = \frac{\text{prop. de verdaderos positivos}}{\text{prop. de falsos positivos}} = \frac{0.935}{0.123} = 7.60$$

$$\text{RV} - = \frac{1 - \text{sensibilidad}}{\text{especificidad}} = \frac{\text{prop. de falsos positivos}}{\text{prop. de verdaderos negativos}} = \frac{0.065}{0.877} = 0.07$$

Ahora ejemplificaremos la determinación del cambio de la probabilidad preprueba y posprueba. De acuerdo con el artículo de Smith AF en este hospital de tercer nivel de atención, entre los pacientes que llegan con sospecha clínica, 64% tiene, verdaderamente, el infarto de miocardio. Supóngase que al atender a un paciente el médico tratante tuvo la impresión de que fue un típico caso que se envía al servicio, por lo que dejó su probabilidad preprueba de tener infarto de miocardio igual que la prevalencia en su servicio, que es de 64%. Al aplicar la prueba de CK resultó positivo con este paciente. Para evaluar cómo repercutió este resultado en la probabilidad posprueba, primeramente hay que convertir la probabilidad preprueba en el momio preprueba. **Figura 3**

$$\text{Momio} = \frac{\text{Probabilidad de preprueba}}{1 - \text{Probabilidad de preprueba}} = \frac{0.64}{1 - 0.64} = 1.78$$

Ahora, se multiplica este momio preprueba por RV+ de la prueba de CK para obtener el momio posprueba. Al resultar como positivo, con esta prueba se aplica como ponderador RV+, así que:

$$\text{Momio posprueba} = \text{Momio preprueba} \times \text{RV+} = 1.78 \times 7.60 = 13.53$$

Finalmente convierte el momio posprueba en la probabilidad posprueba:

$$\text{Probabilidad posprueba} = \frac{\text{Momio posprueba}}{1 + \text{Momio posprueba}} = \frac{13.53}{1 + 13.53} = 0.93$$

Por ende, en este ejemplo, la probabilidad preprueba de 64% se elevó a la probabilidad posprueba de 93% en tener infarto de miocardio. Es muy razonable internar a este paciente en la unidad de cuidados coronarios con base en este resultado.

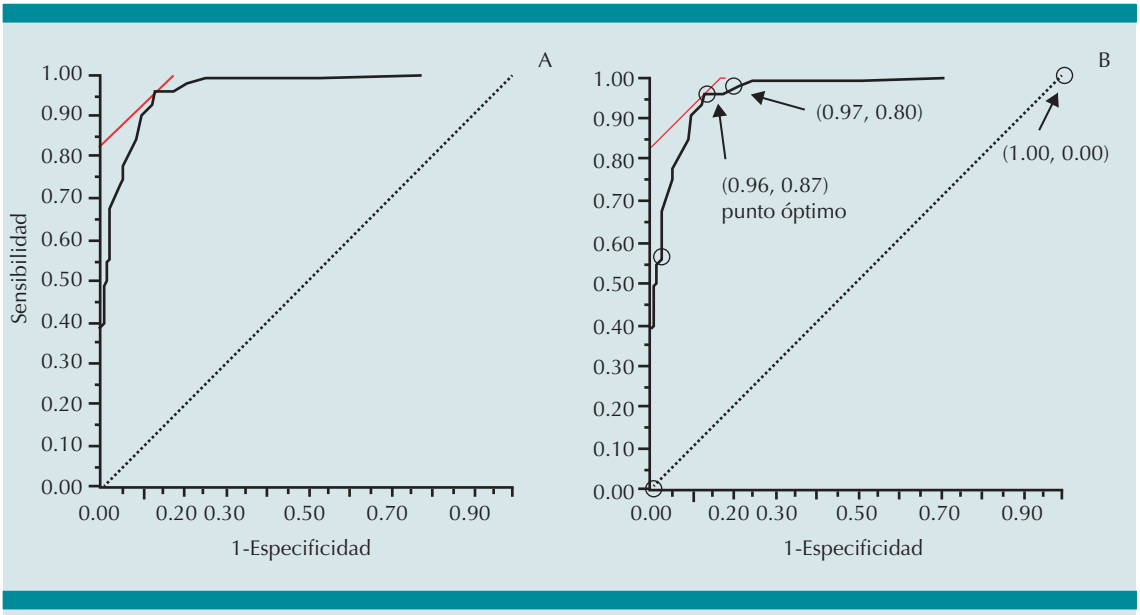

Figura 3. A) Curva ROC construido con los datos de Smith AF, 1967. **B)** Puntos elegidos arbitrariamente. Dentro de los paréntesis se presentaron la sensibilidad y la especificidad. El punto que maximiza el índice *J* corresponde al valor de CK = 72 UI/L. Los estimadores de parámetros calculados utilizando este valor como el punto de corte para el diagnóstico son: AUC=0.96; sensibilidad=0.96; especificidad=0.87; VPP=0.93; VPN=0.92.

¿Qué sucedería si la prueba fuera negativa bajo la misma circunstancia? En este caso, hay que multiplicar el momio preprueba por RV-, 0.07.

$$\text{Momio posprueba} = \text{Momio preprueba} \times RV- = 1.78 \times 0.07 = 0.12$$

Al convertir este momio posprueba en la probabilidad posprueba:

$$\text{Probabilidad posprueba} = \frac{\text{Momio posprueba}}{1 + \text{Momio posprueba}} = \frac{0.12}{1 + 0.12} = 0.11$$

La probabilidad de que este paciente tenga la manifestación clínica que les hizo sospechar el infarto de miocardio a los médicos disminuye de 64 a 11%. En este caso sería bueno observar la evolución del paciente sin internarlo en una unidad de cuidados coronarios.

REFERENCIAS

1. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002; 324: 539-541
2. Haynes RB, You JJ. The architecture of diagnostic research. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis. Theory and methods of diagnostic research, 2nd Ed. OX, UK: BMJ Books, 2009, p.20-41
3. Jaeschke R, Guyatt G, Sackett D.L. Users' Guides to the Medical Literature. III. How to use an article about a diagnostic test. Are the results of the study valid? JAMA 1994; 271: 389-391.
4. Knottnerus JA, Buntinx F, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis. Theory and methods of diagnostic research, 2nd Ed. OX, UK: BMJ Books, 2009, p.1-19
5. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis. Theory and methods of diagnostic research, 2nd Ed. OX, UK: BMJ Books, 2009, p.42-62
6. Irwig LM, Bossuyt PMM, Glasziou PP, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis. Theory and methods of diagnostic research, 2nd Ed. OX, UK: BMJ Books, 2009, p.96-117
7. Smith AF. Diagnostic value of serum-creatine-kinase in a Coronary-Care Unit. Lancet 1967;2(7508):178-182
8. U.S. Centers for Disease Control and Prevention. Principles of epidemiology in public health practice. In: An introduction to applied epidemiology and biostatistics. 3rd ed. Self Study Course SS1000, Lesson 3. https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section1.html Accessed 11 July 2019.
9. Habbema JF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis. Theory and methods of diagnostic research, 2nd Ed. OX, UK: BMJ Books, 2009, p.118-145
10. Wilson MC, Henderson MC, Smetana GW. Chapter 5. Evidence-Based Clinical Decision Making. In: Henderson MC, Tierney LM Jr, Smetana GW, editors. The Patient History. An Evidence-Based Approach to Differential Diagnosis, 2nd Ed. NY, USA: LANGE medical book, McGraw Hill, 2012, p.27-32
11. Simel DL, Sama GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 1991;44(8):763-770
12. A dictionary of epidemiology. Sixth ed. Oxford, UK. International Epidemiological Association, Inc. Oxford University Press; 2014. Parameter; p209
13. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? JAMA 1990;263(2):275-278
14. Obuchowski NA. Sample size calculations in studies of test accuracy. Stat Meth Med Res 1998;7:371-392
15. Flahault A, Cadilhac M, Thomas G. Sample size should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol 2005;58:859–62
16. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006;332(7550):1127-9
17. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193-204
18. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11):e012799. doi: 10.1136/bmjopen-2016-012799.