Importancia del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en el uso de las pruebas de diagnóstico clínico y de laboratorio

Importance of calculation of sensitivity, specificity, and other statistical parameters in the use of clinical and laboratory diagnostic tests

Gilberto J. Vizcaíno-Salazar PhD[1]

Resumen: *la elección de una prueba a realizar para un paciente, así como su interpretación, es un escenario diario al cual el médico se debe enfrentar y para el cual debe aplicar su juicio crítico basado en las evidencias informadas. Es común que cuando se habla de una prueba de diagnóstico clínico o de laboratorio se describan parámetros como la sensibilidad, la especificidad y los valores predictivos positivos y negativos. Estos reflejan las características de una prueba diagnóstica y sirven para decidir en qué momento se deben utilizar (sensibilidad y especificidad de una prueba) o qué significado tiene el resultado de una prueba en un paciente en particular. Cuando se trata de comparar estos parámetros en diferentes pruebas y optar por la que es de mayor utilidad en el diagnóstico de una enfermedad determinada, es indispensable que el médico conozca y aprenda cómo se obtienen estas medidas y cuál es su interpretación para decidir la conducta más apropiada para el paciente. El objetivo de la presente revisión es ofrecer los conceptos estadísticos básicos y simples para la comprensión y aplicación de las pruebas de diagnóstico clínico y de laboratorio.*

Palabras clave: *sensibilidad y especificidad, valor predictivo de las pruebas, curva ROC, funciones de verosimilitud, probabilidad pre y posprueba, prevalencia.*

Vizcaíno-Salazar GJ. *Importancia del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en el uso de las pruebas de diagnóstico clínico y de laboratorio. Medicina & Laboratorio 2017; 23: 365-386.*

[1] *Médico, especialista en Hematología, PhD en Ciencias Médicas. Profesor titular emérito e investigador adscrito al Instituto de Investigaciones Clínicas "Dr. Américo Negrette", Facultad de Medicina, Universidad del Zulia. Maracaibo, Venezuela.*
Correo electrónico: gilvizcaino@gmail.com
Conflicto de intereses: el autor declara que no tiene conflicto de intereses
Medicina & Laboratorio 2017; 23: 365-386
Módulo 28 (Artículos de Reflexión), número 4. Editora Médica Colombiana S.A. 2017©
Recibido el 27 de julio de 2017; aceptado el 30 de agosto de 2017

Como sucede con los elementos de una historia clínica y del examen físico, cada procedimiento de diagnóstico clínico o prueba de laboratorio reúne una serie de características que reflejan la información esperada acerca de un paciente del cual el médico desea saber si posee o no determinada enfermedad. En el presente manuscrito analizaremos varios puntos necesarios para comprender la importancia y aplicabilidad del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en la práctica clínica diaria para determinar: a) qué tan eficientes pueden ser las pruebas de diagnóstico clínico y de laboratorio para clasificar a un sujeto como sano o enfermo, de acuerdo con su real estado de salud; es decir, ¿cuál es el desempeño operativo de la prueba (sensibilidad y especificidad)?, b) cuál es la confiabilidad de la prueba o la reproducibilidad de los resultados, por ejemplo, al ser nuevamente aplicada por otro sujeto, por el mismo sujeto o al compararla con otra prueba que no es usada como estándar de referencia y c) cómo verificar qué tan de acuerdo están dos observadores frente a un fenómeno.

Para esto se presentan las definiciones de los principales parámetros de evaluación de una prueba de diagnóstico clínico y de laboratorio, entre ellos la sensibilidad, la especificidad, los valores predictivos positivo y negativo, las razones de verosimilitud, las probabilidades pre y posprueba, entre otros; además de los procedimientos para su cálculo y las interpretaciones de los resultados. Finalmente, se presentan algunos ejemplos para su mejor comprensión.

Definición de términos

A partir de la necesidad de conocer si un paciente realmente posee o no determinada enfermedad se establecen dos preguntas fundamentales: a) si la enfermedad está presente ¿cuál es la probabilidad de que el resultado de la prueba diagnóstica sea positivo? y b) si la enfermedad no está presente ¿cuál es la probabilidad de que el resultado sea normal o negativo? La primera pregunta define lo que se conoce como **sensibilidad** de una prueba diagnóstica y la segunda incluye el concepto de **especificidad** de una prueba diagnóstica. La definición precisa de estos conceptos sería [1]:

- **Sensibilidad:** proporción de individuos enfermos que poseen una prueba positiva.

- **Especificidad:** proporción de individuos sin la enfermedad que poseen una prueba negativa o normal.

De estos dos conceptos se desprenden, a su vez, dos reglas nemotécnicas con los siguientes acrónimos [1,2]:

1. **SENDES:** cuando un signo, síntoma o prueba diagnóstica posee una alta **SE**nsibilidad (mayor del 95%), obtener un resultado **N**egativo o normal **DES**carta el diagnóstico. Por ejemplo, la pérdida de la pulsación de la vena retiniana se presenta en el 100% de los casos con diagnóstico de presión intracraneal elevada; por lo tanto, si una persona presenta una pulsación normal de la vena retiniana se descarta la presencia de presión intracraneal elevada. Otro ejemplo clínico sería: si la sensibilidad de la historia de edema del tobillo en el diagnóstico de la ascitis es del 92%, en consecuencia, si una persona no tiene historia de edema de tobillo es altamente improbable que posea ascitis.

2. **ESPIN:** cuando un signo o prueba posee una alta **ES**pecificidad (mayor o igual que 95%), un resultado **P**ositivo **IN**dica o confirma el diagnóstico. Por ejemplo, el examen

del tacto rectal tiene una especificidad del 82% para el diagnóstico de cáncer de próstata; por lo tanto, el hallazgo de un nódulo prostático al tacto rectal tiene una alta probabilidad de confirmar el diagnóstico de este tipo de cáncer.

En razón a lo anteriormente expuesto, el conocimiento de las características de una prueba determinada no permite *per se* tener una interpretación exacta del resultado, solo nos dice qué proporción de pacientes, con y sin la enfermedad, podrían presentar un resultado positivo o negativo, respectivamente [3]. En este sentido, el interés del médico es determinar la presencia o ausencia de la enfermedad; en consecuencia, surgen las siguientes preguntas: a) si el resultado es positivo ¿cuál es la probabilidad de que la enfermedad esté presente? y b) si el resultado es negativo ¿cuál es la probabilidad de que la enfermedad no esté presente? La primera pregunta refleja lo que se conoce como **valor predictivo positivo (VPP)** de una prueba diagnóstica y la segunda está relacionada con lo que se denomina **valor predictivo negativo (VPN)** del resultado de una prueba diagnóstica. La definición precisa de estos conceptos sería [1]:

- **Valor predictivo positivo (VPP):** proporción de individuos con una prueba positiva que presentan la enfermedad.
- **Valor predictivo negativo (VPN):** proporción de sujetos con una prueba negativa que no presentan la enfermedad.

De acuerdo con esto se puede inferir que la sensibilidad y la especificidad representan la validez de una prueba diagnóstica, y que el valor predictivo positivo y valor predictivo negativo representan la seguridad de una prueba diagnóstica [4]. En este sentido, podemos calificar una prueba diagnóstica en los parámetros mencionados como excelente (mayor o igual al 95%), buena (entre 80% y 94%), regular (entre 50% y 79%) y mala (menor del 50%) [4].

Cálculo de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo

La determinación de los parámetros de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo se realiza construyendo, en primer lugar, una tabla binaria o de 2x2, a partir de la cual se obtienen los diversos resultados probables de una prueba diagnóstica, como se esquematiza en la **figura 1** [1,3-5].

Los posibles resultados de una prueba diagnóstica se interpretan de la manera siguiente [1,3-5]:

	Enfermedad presente	**Enfermedad ausente**
Prueba positiva	Verdaderos positivos (VP) a	Falsos positivos (FP) b
Prueba negativa	Falsos negativos (FN) c	Verdaderos negativos (VN) d

Figura 1. Tabla de 2x2 de una prueba diagnóstica y resultados probables de la misma.

- **Verdaderos positivos:** la enfermedad está presente y se diagnostica al paciente como enfermo.

- **Falsos positivos**: el paciente no presenta la enfermedad y se le diagnostica como enfermo.

- **Verdaderos negativos:** la enfermedad no está presente y se diagnostica al paciente como sano.

- **Falsos negativos**: la enfermedad está presente y se diagnostica al paciente como sano.

Posteriormente, para el cálculo de los parámetros anteriormente mencionados, se toman los datos registrados en la tabla de 2x2 y se aplican las fórmulas presentadas en la **tabla 1** [1,3-5].

Como regla nemotécnica, las fórmulas expresadas anteriormente se pueden resumir en la representación gráfica de la **figura 2**.

Una de las ventajas de los parámetros de sensibilidad y especificidad es que al ser calculados en forma «vertical», dentro de la matriz de decisiones (tabla 2x2), no dependen de la prevalencia de la enfermedad, como sí ocurre con los valores predictivos positivos y negativos que son calculados en forma horizontal (véase **figura 2**) [1,3-5].

Ejemplos para el cálculo de los parámetros de una prueba diagnóstica

Veamos el siguiente ejemplo:

En un estudio publicado [6] los resultados de la prueba de tolerancia al ejercicio (del inglés, *Tread mill exercise*) fueron comparados entre sujetos con y sin enfermedad coronaria. Los criterios usados para el diagnóstico de la enfermedad coronaria fueron la presencia de al menos 70% de estrechez de una o más arterias coronarias, determinada por angiografía coronaria. Una prueba de ejercicio positiva fue definida como el hallazgo de más de un milímetro de aplanamiento o cambios en el segmento ST mayor de 0,08 segundos en comparación con el trazado basal. Se estudió un total de 623 sujetos, de los cuales se encontró evidencia de enfermedad coronaria en 418; de estos últimos 328 tuvieron una prueba de esfuerzo positiva. De los 205 sujetos que no tuvieron evidencia de enfermedad 24 presentaron una prueba positiva.

Ahora, analicemos esquemáticamente este ejemplo consignando los datos en una tabla de 2x2 (véase **figura 3**).

A partir de los datos de la tabla de 2x2 (véase **figura 3**) calculamos los parámetros de las pruebas diagnósticas aplicando las fórmulas consignadas en la **tabla 1**.

De acuerdo con estos resultados, es posible concluir que la prueba de tolerancia al ejercicio posee una sensibilidad moderada (79%), por lo que un resultado negativo no descarta la presencia de la enfermedad coronaria (no se cumple el acrónimo SENDES), pues un 21% de los sujetos tuvieron resultados falsos negativos con esta prueba, pero poseían la enfermedad coronaria demostrada por angiografía. Del mismo modo, la prueba mostró una especificidad del 88% que, aunque es buena, no confirma el diagnóstico si la prueba es positiva (no se cumple el acrónimo ESPIN), pues el 12% de los sujetos presentaron resultados falsos positivos. Ahora bien, el valor predictivo positivo del 93% revela que una prueba positiva asegura con mucha probabilidad el diagnóstico

Tabla 1. Fórmulas para el cálculo de parámetros de una prueba diagnóstica

Sensibilidad=a/(a+c)	que es lo mismo que	*Sensibilidad=VP/(VP+FN)*
Especificidad=d/(b+d)	que es igual que	*Especificidad=VN/(FP+VN)*
Valor predictivo positivo=a/(a+b)	*o*	*Valor predictivo positivo=VP/(VP+FP)*
Valor predictivo negativo=d/(c+d)	*o*	*Valor predictivo negativo=FN/(FN+VN)*
% de falsos negativos=100-sensibilidad		
% de falsos positivos=100-especificidad		
Exactitud=(a+d) / n	*o*	*Exactitud=(VP+VN) / n*

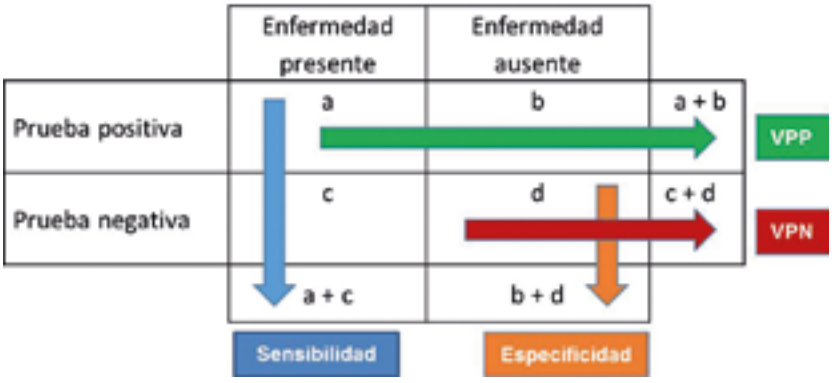

Figura 2. Representación esquemática direccional de las fórmulas para calcular la sensibilidad, la especificidad, el valor predictivo positivo (VPP) y el valor predictivo negativo (VPN). Nótese que cada flecha comienza con la letra dividendo y termina con el divisor (p. ej. a / (a + c) = sensibilidad).

	Enfermedad coronaria presente	**Enfermedad coronaria ausente**	
Prueba de tolerancia al ejercicio positiva	328 a	24 b	352 a + b
Prueba de tolerancia al ejercicio negativa	90 c	181 d	271 c + d
	418 a + c	205 b + d	623 a+b+c+d

Figura 3. Tabla 2x2 ejemplo 1.

de la enfermedad coronaria, mientras que un valor predictivo negativo del 67% indica que la prueba no excluye la enfermedad coronaria a pesar de que haya resultado negativa.

Examinemos otro ejemplo:

En un estudio realizado [7] para evaluar la respuesta a la insulina en pacientes obesos y no obesos (de acuerdo con los parámetros del índice de masa corporal) y determinar la posibilidad de encontrar resistencia a la insulina, como signo presuntivo para desarrollar diabetes mellitus tipo 2 en individuos con valores de glicemia menores a 110 mg/dL, se incluyeron un total de 135 sujetos, 58 de sexo masculino y 77 de sexo femenino. En los pacientes de sexo masculino se encontraron 38 obesos, 31 con respuesta insulínica anormal y siete con respuesta normal, y 20 no obesos, uno con respuesta insulínica anormal y 19 con respuesta normal. Veamos la representación esquemática para examinar la prueba de tolerancia a la insulina en el sexo masculino.

Nuevamente, a partir de los datos de la tabla de 2x2 (véase **figura 4**) calculamos los parámetros de las pruebas diagnósticas aplicando las fórmulas consignadas en la **tabla 1**.

De estos resultados se deduce que la prueba de tolerancia a la insulina (o de resistencia a la insulina) posee una alta posibilidad de predecir el desarrollo a futuro de la diabetes mellitus tipo 2, en individuos obesos del sexo masculino. A pesar de su buena sensibilidad (82%), un resultado normal no descarta la posibilidad de desarrollar diabetes mellitus tipo 2 (no se cumple el SENDES), pero ante su alta especificidad (95%) un resultado positivo indica que es casi segura la posibilidad de presentar diabetes mellitus tipo 2 (se cumple el acrónimo ESPIN). Asimismo, en individuos con prueba positiva (valor predictivo positivo) existe una muy alta probabilidad (97%) de que aquellos de sexo masculino y obesos desarrollen diabetes mellitus tipo 2; pero ante una prueba negativa (valor predictivo negativo) no se puede excluir el diagnóstico de diabetes mellitus tipo 2, especialmente en individuos obesos del sexo masculino. En conclusión, la relación importante entre la obesidad y la presencia de diabetes mellitus tipo 2 se puede determinar con la realización de una prueba de resistencia a la insulina en la población de sexo masculino.

	Individuos obesos	Individuos no obesos	
Respuesta insulínica anormal	31 a	1 b	32 a + b
Respuesta insulínica normal	7 c	19 d	26 c + d
	38 a + c	20 b + d	58 a+b+c+d

Figura 4. Tabla 2x2 ejemplo 2.

Curva ROC en términos de la sensibilidad y la especificidad

La curva ROC (del inglés, *Receiver Operating Characteristic*), o curva característica de funcionamiento del receptor, fue desarrollada por ingenieros eléctricos para medir la eficacia en la detección de objetos enemigos en campos de batalla mediante pantallas de radar, a partir de lo cual se elaboró la teoría de detección de señales (TDS). El análisis ROC se aplicó posteriormente en medicina, radiología, psicología y otras áreas durante varias décadas. En la curva ROC se presenta la sensibilidad de una prueba diagnóstica, que produce resultados continuos en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte [8].

Por su parte, la separación de los grupos, con y sin enfermedad, representa la capacidad discriminatoria de una prueba para clasificar a los sanos como sanos y a los enfermos como enfermos. Un parámetro para evaluar la bondad de una prueba diagnóstica, que produce resultados continuos, es el área bajo la curva (AUC; del inglés, *Area Under the Curve*), la cual se puede interpretar como la probabilidad de que ante un par de individuos, uno enfermo y el otro sano, la prueba los clasifique correctamente [8].

La realidad nos indica que una amplia gama de pruebas diagnósticas reportan sus resultados cuantitativamente, utilizando escalas continuas (p. ej. recuento de leucocitos, proteína C reactiva hipersensible, homocisteína plasmática, nivel de glucosa en sangre, entre otras). En estas pruebas, para establecer el diagnóstico de una determinada enfermedad, se establece un punto de corte (*cut off*) por encima del cual se apoya el diagnóstico y por debajo del cual se rechaza, o viceversa [8-10].

El análisis con base en las curvas ROC constituye un método estadístico para determinar la exactitud diagnóstica de las pruebas que utilizan escalas continuas con tres propósitos específicos: a) determinar el punto de corte en el que se alcanza la sensibilidad y la especificidad más alta, b) evaluar la capacidad discriminativa de la prueba diagnóstica, es decir, su capacidad de diferenciar sujetos sanos frente enfermos y c) comparar la capacidad discriminativa de dos o más pruebas diagnósticas que expresan sus resultados como escalas continuas [9-11].

Los ejes del gráfico de curva ROC adoptan valores entre 0 y 1 (0% y 100%), lo que delimita un cuadrado de área igual a 1,00. Una prueba diagnóstica se considera no discriminativa cuando su curva ROC coincide con la línea de no discriminación, la cual posee un área bajo la curva de 0,50 (véase **figura 5**). A medida que el AUC de una prueba diagnóstica se acerca al valor 1,00 (prueba diagnóstica perfecta) mayor será su capacidad discriminativa [9].

El punto de corte de una escala continua, que determina la sensibilidad y especificidad más alta, es aquel que presenta el mayor índice de Youden, calculado según la fórmula [sensibilidad + (1 - especificidad)]. Gráficamente, este corresponde al punto de la curva ROC más cercano al ángulo superior izquierdo del gráfico (punto 1,0); es decir, más cercano al punto del gráfico cuya sensibilidad y especificidad es igual a 100% (véase **figura 5**). Es preciso aclarar que el índice de Youden identifica el punto de corte que determina la sensibilidad y especificidad más alta conjuntamente, es decir, para un

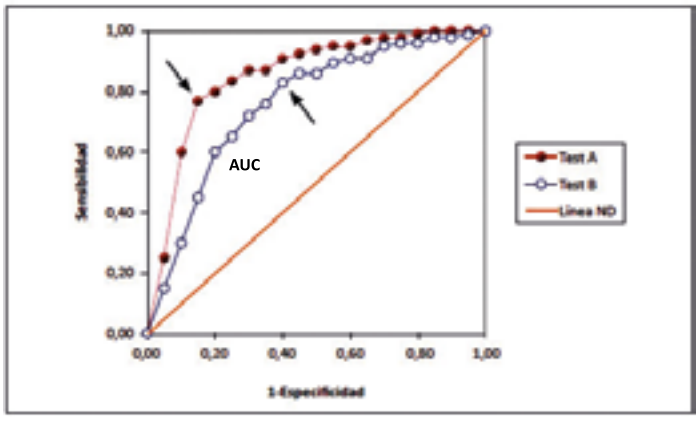

Figura 5. Representación gráfica de curvas ROC de dos pruebas diagnósticas hipotéticas (A y B). Para cada curva ROC las flechas indican el punto de corte que determina la sensibilidad y la tasa de falsos positivos (1- especificidad) conjunta más alta (índice de Youden). AUC denota el área bajo la curva. La línea recta, en naranja, representa la línea de no discriminación (línea ND), la cual divide en dos mitades iguales el cuadrado de área de 1,00, por lo que su AUC es igual a 0,50.

mismo punto; sin embargo, dicho punto de corte no necesariamente determina la sensibilidad ni la especificidad más alta que podría alcanzar la prueba, ya que, generalmente, la sensibilidad más alta es determinada por un punto de corte, mientras que la especificidad más alta es determinada por otro [9].

Para comprender mejor el concepto de discriminación es más simple pensar que el eje Y del gráfico de la curva ROC corresponde a la proporción de verdaderos positivos sobre el total de pacientes enfermos, es decir, la sensibilidad, y que el eje X corresponde a la proporción de falsos positivos sobre el total de sujetos sanos, es decir, 1 - especificidad. Visto de esta manera, un gráfico de curva ROC ilustra la «proporción de verdaderos positivos» (eje Y) versus la «proporción de falsos positivos» (eje X) para cada punto de corte de una prueba diagnóstica cuya escala de medición es continua [8].

Para una construcción e interpretación correcta de los gráficos de curvas ROC es necesario tener en cuenta tres conceptos finales. En primer lugar, la sensibilidad, la especificidad y el área bajo la curva son estimadores muestrales de parámetros poblacionales; por consiguiente, cada uno tiene asociado un error de estimación, lo que hace necesario reportar sus respectivos intervalos de confianza (IC). En segunda instancia, los estudios de exactitud diagnóstica, a partir de los cuales se construyen las curvas ROC, corresponden generalmente a diseños de tipo transversal, es decir, de casos y controles. Finalmente, la validez de estos estudios (p. ej. el riesgo de sesgo) debe ser evaluada críticamente, de acuerdo con las recomendaciones descritas por la medicina basada en la evidencia [12].

De esta manera, las curvas ROC son útiles para:

1. Conocer el rendimiento global de una prueba (área bajo la curva o AUC).

2. Comparar dos o más pruebas o dos puntos de corte de una misma entidad nosoló-

gica: comparación de dos curvas o dos puntos sobre una curva.

3. Elegir el punto de corte apropiado para un determinado paciente.

Entre las limitaciones de su uso se encuentran: solo contemplan dos estados clínicos posibles (sano, enfermo) y no sirven para situaciones en que se trata de discernir entre más de dos enfermedades.

Ilustremos un ejemplo hipotético sobre la comparación de tres pruebas mediante la construcción de curvas ROC. Considérense 100 resultados positivos y 100 resultados negativos presentados en la **figura 6**.

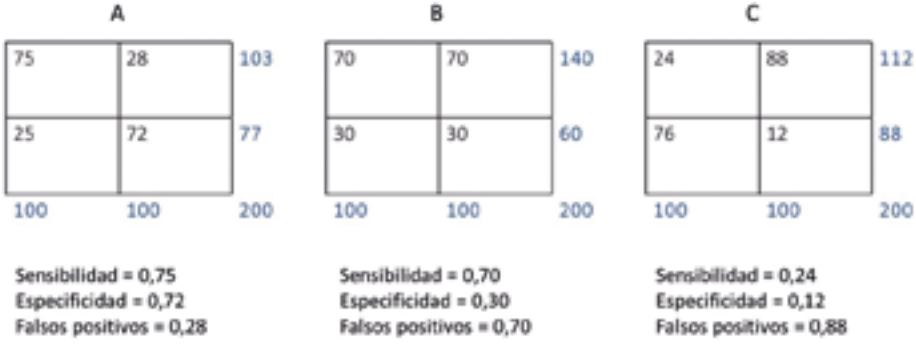

Figura 6. Resultados de tres pruebas (A, B, C).

En la **figura 7** se grafican, en el espacio ROC, los puntos de los tres ejemplos anteriores. El resultado del método A muestra claramente que es el mejor entre los métodos A, B y C. El resultado del método B se observa sobre la línea de no discriminación (diagonal). Si aplicamos además la fórmula para calcular la exactitud vemos que es de 73,5%, 50,0% y 28,0% para los métodos A, B y C, respectivamente. El método C aparece como el peor de los tres, con un resultado muy pobre. Además, se considera que el método A tiene mejor rendimiento porque su área bajo la curva tiene un punto de inflexión que se acerca a la prueba perfecta (1,0).

Dado que el interés de esta revisión es presentar al profesional de la salud una metodología simple, sencilla y de fácil comprensión al lector, no se ahondará en los cálculos matemáticos para las curvas ROC, los cuales ameritan una mayor explicación. En este caso se sugiere consultar otras referencias que abordan el tema con mayor profundidad [13-15].

Razón de verosimilitud

La razón de verosimilitud es la probabilidad de que el resultado de una prueba determinada se espere en un paciente con una enfermedad concreta (razón de verosimilitud positiva), comparado con la probabilidad de que el mismo resultado se espere en un paciente con otra enfermedad (razón de verosimilitud negativa) [4]. Por ejemplo, si se

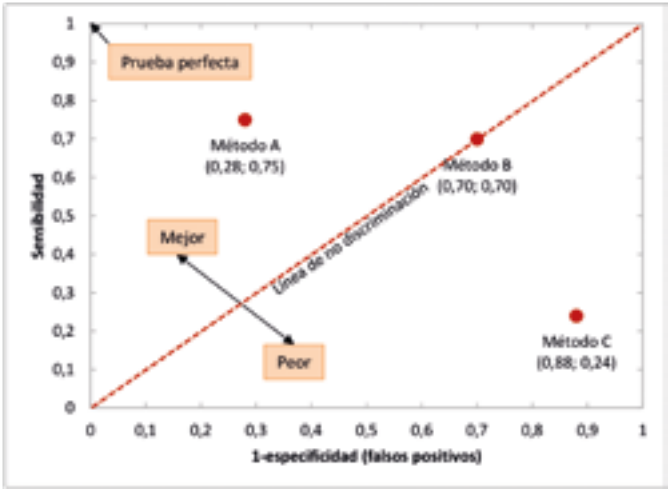

Figura 7. Representación gráfica mediante elaboración de curvas ROC de los datos de la **figura 6**.

tiene un paciente con anemia y una ferritina sérica de 18 ng/mL (valor normal = 30 ng/mL a 400 ng/mL) y en un artículo se encuentra que el 90% de los pacientes con anemia por deficiencia de hierro poseen niveles de ferritina sérica en el mismo rango que este paciente (sensibilidad), pero que existe un 15% de los pacientes con otras causas de anemia que tienen ferritina sérica con el mismo valor que el paciente en mención (1 - especificidad), entonces el resultado de dicho paciente podría ser seis veces más probable (90/15) de ser diagnosticado con anemia ferropénica que con anemia por otras causas; esto se denomina razón de verosimilitud para un resultado positivo de una prueba.

Cálculo de la razón de verosimilitud positiva y negativa

Razón de verosimilitud positiva
=sensibilidad/(1-especificidad)

La razón de verosimilitud positiva se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos entre la probabilidad de un resultado positivo en los sanos. Es, en definitiva, el cociente entre la fracción de verdaderos positivos (sensibilidad) y la fracción de falsos positivos (1-especificidad) [1].

Razón de verosimilitud negativa
=(1-sensibilidad)/especificidad

La razón de verosimilitud negativa se calcula dividiendo la probabilidad de un resultado negativo en presencia de la enfermedad entre la probabilidad de un resultado negativo en ausencia de la misma. Se calcula, por lo tanto, como el cociente entre la fracción de falsos negativos (1-sensibilidad) y la fracción de verdaderos negativos (especificidad) [1].

Pre y posanálisis de los parámetros de las pruebas diagnósticas y de laboratorio

Existen situaciones donde se desea comparar las bondades de una prueba diagnóstica

antes y después de un tratamiento específico para determinada enfermedad. En un estudio realizado por Gisbert y colaboradores (2003) [16], en el que utilizaron la prueba de aliento con ^{13}C-urea para el diagnóstico de la infección con *Helicobacter pylori,* encontraron en el grupo pretratamiento los siguientes resultados: sensibilidad= 96% (IC 95% = 81% a 99%), especificidad = 100% (IC 95% = 69% a 100%), valor predictivo positivo = 100% (IC 95% = 87% a 100%), valor predictivo negativo = 92% (IC 95% = 59% a 100%), razón de verosimilitud positiva = infinito, y razón de verosimilitud negativa = 0,04. Luego del tratamiento los resultados fueron los siguientes: sensibilidad= 100% (IC 95% = 77% a 100%), especificidad = 97% (IC 95% = 90% a 99%), valor predictivo positivo = 88% (IC 95% = 62% a 98%), valor predictivo negativo = 100% (IC 95% = 95% a 100%), razón de verosimilitud positiva = 35 y razón de verosimilitud negativa = 0, lo que confirma la excelencia de esta prueba.

Probabilidad preprueba

La probabilidad preprueba es definida como la probabilidad de presentar una enfermedad determinada antes de que el resultado de una prueba sea conocido, basada en las manifestaciones clínicas del paciente. Dicha probabilidad preprueba para pacientes con sospecha de un diagnóstico determinado (PD^+) se puede interpretar como la proporción de pacientes con la enfermedad (D^+), entre todos los pacientes que presentan un síntoma o conjunto de síntomas ($D^+ + D^-$), siendo la fórmula la siguiente:

$$PD^+ = D^+ / (D^+ + D^-)$$

Donde (D^+) es el número de pacientes con la enfermedad, (D^-) es el número de pacientes sin la enfermedad y PD^+ es la probabilidad de presentar la enfermedad.

Como ejemplo, un grupo de investigadores norteamericanos [17] estudiaron las enfermedades que presentan los pacientes con mareos persistentes. De un total de 100 pacientes con mareos 16 tenían una condición psiquiátrica; se deduce entonces que 16% es el estimado para la probabilidad de presentar una condición psiquiátrica en es te estudio. Por lo tanto, se puede inferir que si uno de sus pacientes presenta mareos persistentes la probabilidad (PD^+) de presentar una enfermedad psiquiátrica es del 16%.

La probabilidad preprueba es especialmente útil en cuatro situaciones:

1. En la interpretación de los resultados de una prueba diagnóstica.

2. En la selección de una o más pruebas diagnósticas.

3. En la decisión de comenzar el tratamiento:

- Sin pruebas adicionales

- Mientras se espera el resultado de una prueba

4. En la decisión si la prueba tiene importancia.

Todo paciente en el que se sospecha una enfermedad tendrá una probabilidad de presentarla, la cual dependerá de la **prevalencia** de dicha condición clínica en la población, las características del paciente (p. ej. edad, género, raza), los signos y síntomas presentes, entre otros. Así, antes de realizar cualquier prueba diagnóstica el clínico, explícita o implícitamente, le asigna a su paciente una **probabilidad preprueba** de presentar la enfermedad que está determinada por la prevalencia de dicha afección. Una vez realizada la prueba diagnóstica, la probabilidad de presentar la enfermedad

aumentará o disminuirá, dependiendo del resultado de la prueba en cuestión. Esta nueva probabilidad se llama **probabilidad posprueba** [18].

Una forma de aproximarse a la probabilidad preprueba en un paciente determinado es utilizar la prevalencia de la enfermedad en el estudio que se está analizando (total de pacientes positivos para la enfermedad, o (a + c) en nuestra tabla de contingencia, sobre el total de pacientes del estudio, o (a + b + c + d). Si nuestro paciente es similar a la población del estudio sería razonable utilizar este valor.

La probabilidad preprueba de la enfermedad es la probabilidad de que la afección del paciente objetivo esté condicionada a la información clínica antes de realizar la prueba diagnóstica. La probabilidad preprueba se puede calcular sobre la base del juicio clínico, la prevalencia media de la población a la que pertenece el paciente o variables clínicas, con o sin información de los resultados de las pruebas diagnósticas. Por ejemplo, los modelos de predicción multivariables para la presencia de enfermedad coronaria aterosclerótica se pueden utilizar para calcular la probabilidad preprueba de esta enfermedad, con base en variables clínicas específicas del paciente [19].

La probabilidad preprueba afecta el valor predictivo positivo y el valor predictivo negativo de una prueba diagnóstica y, por lo tanto, determina la utilidad, por ejemplo, de una prueba de imagen. La angiografía coronaria computarizada es una excelente prueba de imagen para el diagnóstico de la enfermedad coronaria aterosclerótica, con una sensibilidad del 98% y una especificidad del 89% respecto a la angiografía coronaria invasiva [20]. En este caso, cuando la probabilidad preprueba para una angiografía coronaria computarizada es del 20%, el valor predictivo positivo es del 69%, mientras que cuando la probabilidad preprueba es del 80% el valor predictivo positivo es del 97%.

Diferencias entre razón de verosimilitud y probabilidad preprueba

Cuando se calcula la razón de verosimilitud se tiene en cuenta, en primer lugar, una condición patológica conocida, los síntomas y signos asociados y el tipo o características de los pacientes. Por el contrario, al emplear la probabilidad preprueba se considera, en primera instancia, la sintomatología que presenta el paciente y, luego, la probabilidad de que el paciente presente una enfermedad determinada. En las **tablas 2** y **3** se presentan ejemplos de estos dos conceptos.

Según los resultados obtenidos es posible evaluar la calidad de las pruebas. De esta manera, una prueba con una razón de verosimilitud positiva mayor que 10 se considera excelente, entre 5 y 10 buena, entre 5 y 2 regular y menor que 2 inútil. Por su parte, aquella prueba con una razón de verosimilitud negativa entre 0,5 y 1,0 se considera inútil, entre 0,2 y 0,5 regular, entre 0,1 y 0,2 bueno y menor que 0,1 excelente [41].

La ventaja de utilizar los parámetros de razón de verosimilitud positiva y negativa de una prueba diagnóstica radica en que no dependen de la proporción de enfermos en la muestra, sino solo de la sensibilidad y especificidad de la prueba; de ahí su utilidad a la hora de comparar pruebas diagnósticas. La probabilidad preprueba suele ser conocida y no es más que la prevalencia de la enfermedad que queremos diagnosticar; además, se puede obtener por medio de

Tabla 2. Ejemplos del cálculo de la razón de verosimilitud en algunas condiciones patológicas

Condición patológica	Síntoma, signo o prueba diagnóstica	Tipo de pacientes	Razón de verosimilitud positiva
Sinusitis [21]	Pacientes con odontalgia maxilar, secreción nasal, pobre respuesta a descongestionantes nasales, transiluminación anormal, historia de secreción nasal purulenta	Con 4 signos o síntomas Con 3 signos o síntomas Con 2 signos o síntomas Con 1 signo o síntoma Sin signos o síntomas	6,4 2,6 1,1 0,5 0,1
Ascitis [22]	Presencia de onda líquida	Sexo masculino	9,6
Insuficiencia cardíaca congestiva [23]	Presencia de reflujo abdomino-yugular	Con disnea	6,0
Hipertensión renovascular [24]	Murmullo sistólico o diastólico abdominal	Hipertensos	39,0
Cirrosis hepática [25]	Ascitis Trombocitopenia Telangectasias Puntaje en la escala de Bonacini para cirrosis hepática mayor que 7	Con hepatomegalia	7,2 6,3 4,3 9,4
Accidente cerebrovascular isquémico [26]	Proteína C reactiva de alta sensibilidad punto de corte mayor que 5,58 mg/L	Con hipertensión arterial	3,9
Anemia por deficiencia de hierro [27]	Ferritina sérica	Paciente con anemia y ferritina mayor que 100 µg/L Ferritina entre 45 µg/L y 100 µg/L Ferritina entre 18 µg/L y 45 µg/L Ferritina menor que 18 µg/L	0,13 0,46 3,12 41,47
Neurocisticercosis [28]	ELISA (a) e inmunoblot (b) para neurocisticercosis	Con epilepsia y en zona endémica	11,3 (a) 19,6 (b)
Tuberculosis intestinal frente enfermedad de Crohn[29]	Reacción en cadena de la polimerasa (PCR) para tuberculosis	Con enfermedad crónica intestinal	10,68
Cirrosis hepática [30]	Determinación de interleuquina-6 (IL-6)	Infección bacteriana en pacientes cirróticos	8,99

Tabla 3. Ejemplos del cálculo de la probabilidad preprueba

Paciente problema	Enfermedad o prueba diagnóstica	Probabilidad preprueba (%)
Pacientes sin enfermedad coronaria conocida, con sospecha de angina inestable, mayor de 50 años y sexo masculino [31]	Enfermedad coronaria demostrable por ecocardiografía de estrés	35
Paciente con dispepsia crónica [32]	Infección por *Helicobacter pylori* demostrable por prueba de aliento con ^{13}C-urea	25
Paciente con dolor, eritema y edema en miembro inferior izquierdo [33]	Trombosis venosa profunda demostrable por ecografía Doppler	75
Paciente con dolor abdominal en hipocondrio derecho [34]	Sospecha de litiasis biliar por ultrasonido abdominal	40,8
Paciente de 75 años, no fumador, no hipertenso, con perfil lipídico normal [35]	Enfermedad cardiovascular arteriosclerótica y determinación de hemoglobina glucosilada mayor de 6,5%	12,4
Paciente con aterosclerosis carotidea [36]	Sospecha de microembolización medida por ecografía Doppler transcraneal	3
Paciente con sospecha de cáncer pancreático [37]	Diagnóstico por tomografía computarizada de tumor irresecable previo laparoscopia	41,4
Recién nacido con síndrome de TORCH [38]	Descarte citomegalovirus por reacción en cadena de la polimerasa (PCR)	4
Paciente con cefalea intensa y rigidez de nuca [39]	Descarte de hemorragia subracnoidea por tomografía computarizada	20
Paciente con nódulo tiroideo palpable [40]	Detección de malignidad por biopsia por aspiración con aguja fina guiada por ultrasonido	5 a 20

una estimación aproximada con base en la experiencia profesional o en datos estadísticos o epidemiológicos de la enfermedad a la cual se le aplique la prueba diagnóstica [41].

Fagan T. J., en 1975 [42], describió un nomograma para el teorema de Bayes basado en la capacidad de convertir el teorema de Bayes en una función sumatoria lineal simple. El nomograma de Fagan (véase **figura 8**) tiene tres columnas: la primera es la probabilidad preprueba de tener la enfermedad antes de aplicar la prueba (prevalencia), la segunda es la razón de verosimilitud y la tercera la probabilidad posprueba. Con una regla se traza una línea entre la probabilidad preprueba y la razón de verosimilitud. La prolongación de esta línea corta en la tercera columna es la probabilidad posprueba de tener la enfermedad en función del resultado de la prueba.

De acuerdo con este nomograma, a través de una probabilidad preprueba conocida se puede calcular la probabilidad posprueba por intermedio de la razón de verosimilitud, uniendo los valores por medio de una línea recta. Una razón de verosimilitud menor que 1 produce una probabilidad

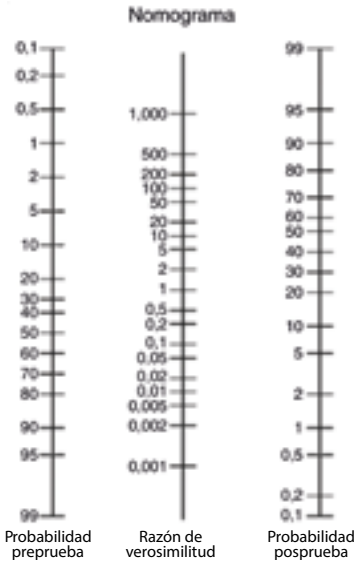

Figura 8. Nomograma de Fagan.

posprueba mucho más baja que la probabilidad preprueba. Cuando la probabilidad preprueba yace entre el 30% y el 70%, las pruebas diagnósticas con una razón de verosimilitud alta (por encima de 10) confirman el diagnóstico. Por el contrario, una razón de verosimilitud baja (menor que 0,1) descarta la probabilidad de que el paciente presente la enfermedad [1].

Influencia de la prevalencia en los resultados de las pruebas diagnósticas

La prevalencia se define como el número total de casos de una población que padecen una enfermedad sobre un tiempo dado, o lo que es lo mismo, el número total de casos de la población con la enfermedad dividido por el número total de individuos de dicha población, que con base en la tabla binaria de 2x2 mostrada anteriormente (véase **figura 1**) se expresaría mediante la siguiente fórmula:

$$Prevalencia = a+c \,/\, a+b+c+d$$

Es importante aclarar que la prevalencia es distinta a la incidencia. La prevalencia es una medición de todos los individuos afectados por una enfermedad en un periodo en particular, mientras que la incidencia es una medición del número de **nuevos** individuos que contraen una enfermedad durante un periodo específico. Si conocemos la prevalencia podemos también calcular lo que denominamos posibilidades (del inglés, *odds*) preprueba y posprueba y calcular, posteriormente, la probabilidad posprueba sin necesidad de usar el nomograma de Fagan, así:

Posibilidades (odds) preprueba
=prevalencia / (1-prevalencia)

Posibilidades (odds) posprueba
= posibilidades (odds) preprueba
x razón de verosimilitud

Probabilidad posprueba
= posibilidades (odds) posprueba/
[posibilidades (odds) posprueba+1]

A continuación, se aplican los parámetros descritos en esta revisión en un ejemplo práctico, en el que se utilizan los resultados de una revisión sistemática sobre la ferritina sérica como prueba diagnóstica para la anemia por deficiencia de hierro (véase **figura 9**) [1].

	Anemia ferropénica presente	Anemia ferropénica ausente	
Ferritina positiva (< 65 mmol/L)	731 a	270 b	1.001 a + b
Ferritina negativa (>65 mmol/L)	78 c	1.500 d	1.578 c + d
	809 a + c	1.770 b + d	2.579 a+b+c+d

Figura 9. Tabla 2x2 ejemplo 3.

Nuevamente, a partir de los datos de la tabla de 2x2 (véase **figura 9**) calculamos los parámetros de las pruebas diagnósticas aplicando las fórmulas revisadas hasta el momento.

Sensibilidad=731/809
= 0,903 x 100 =90,3%

Especificidad
=1.500/1.770=0,847 x 100=84,7%

Valor predictivo positivo
=731/1.001 =0,73 x 100=73%

Valor predictivo negativo
=1.500/1.578 =0,95 x 100=95%

% de falsos negativos
=100-90,3=9,7%

% de falsos positivos
=100-84,7=15,3%

Razón de verosimilitud positiva
=90,3/15,3 =5,90

Razón de verosimilitud negativa
=9,7/84,7 =0,11

Prevalencia=809/2.579
=0,314 x 100= 31,4%

Posibilidades preprueba
=31,3/68,7 =0,45

Posibilidades posprueba
=0,45 x 5,90=2,65

Probabilidad posprueba
=2,65/(2,65+1)=2,65/3,65=0,726 x 100=72,6%

Con base en los resultados anteriores, podemos inferir que la ferritina sérica presenta una alta sensibilidad (90,3%), por lo que ante un resultado negativo es muy probable que se descarte el diagnóstico de anemia ferropénica (aunque no hay SENDES porque hay un 9,7% de falsos negativos) y una especificidad moderada (84,7%), que demuestra que un resultado positivo no indica o confirma la posibilidad de tener anemia ferropénica (no hay ESPIN pues hay un 15,3% de falsos positivos). Por su parte, los resultados de los valores predictivos sugieren que una ferritina positiva no siempre es indicativa de anemia ferropénica (VPP= 73%), pero

una ferritina negativa indica que casi con seguridad el paciente no posee anemia ferropénica (VPN = 95%).

La probabilidad preprueba (prevalencia) también se puede calcular al extrapolar al nomograma de Fagan (véase **figura 8**) los valores de la probabilidad posprueba (72,6%) y la razón de verosimilitud positiva (5,9), y trazar una línea recta que una estos puntos; esto da un valor aproximado de 30%, cercano a lo calculado mediante la fórmula para la prevalencia (31,4%). En conclusión, la prueba de la ferritina sérica es una excelente herramienta para descartar anemia por deficiencia de hierro cuando el resultado es negativo, pero no confirma el diagnóstico a pesar de que el resultado sea positivo.

El hallazgo de que la prevalencia puede influir en las estimaciones de validez de las pruebas diagnósticas ofrece una pista para identificar defectos metodológicos en los estudios de evaluación de dichas pruebas [43].

Consideraciones sobre las pruebas de laboratorio en el diagnóstico clínico

En cada prueba de laboratorio relacionada con el diagnóstico de una enfermedad determinada se deben considerar, además del control de calidad exigido como requisito indispensable para la estimación de los resultados, su confiabilidad, reproducibilidad, validez interna (variaciones intrapruebas) y validez externa (variaciones interpruebas). Así, por ejemplo, en el caso de las lesiones pancreáticas focales es importante caracterizar si son no cancerosas, precancerosas o cancerosas. Para esto, se encuentran disponibles ciertos exámenes que permiten caracterizar la naturaleza de este tipo de lesiones, entre los que se incluyen el examen por tomografía axial computarizada (TAC), la imagenología por resonancia magnética (MRI), la tomografía de emisión de positrones (TEP), la ecografía endoscópica (también conocida como endosonografía o EE), elastografía por ultrasonografía endoscópica (USE) y la biopsia guiada por ecografía endoscópica[44].

En 2017, Best y colaboradores [44] realizaron una búsqueda bibliográfica minuciosa de los estudios que informaban sobre la exactitud de diferentes exámenes imagenológicos para detectar lesiones cancerosas y precancerosas en personas con lesiones pancreáticas focales hasta el 19 de julio de 2016. Luego de aplicar diversos criterios incluyeron 54 estudios que proporcionaban información sobre 3.196 pacientes con este tipo de lesiones evaluadas por imagenología, y se compararon los resultados de las diversas pruebas con el diagnóstico final entregado por el estándar de referencia (examen histopatológico de la lesión obtenida mediante extracción quirúrgica). No obstante, los autores indicaron que no fue posible establecer conclusiones sólidas debido: a) las diferencias en la forma en las que los autores de los estudios originales clasificaron las lesiones pancreáticas focales en lesiones cancerosas, precancerosas y benignas, b) la inclusión de pocos estudios con amplios intervalos de confianza para cada comparación, c) la calidad metodológica deficiente de los estudios y d) la heterogeneidad en los cálculos dentro de las comparaciones.

Con base en lo anterior, tomemos otro ejemplo. Supongamos que se decide incrementar el nivel normal (punto de corte) de la prueba de hemoglobina glucosilada (HbA1c) a 6%, cuando anteriormente era menor de 5,6%. Esto modificaría todos los

parámetros destinados al cálculo de las bondades de esta prueba, pero el valor que más se alteraría sería el valor predictivo negativo porque ahora se consideran como normales los pacientes que presentan valores por debajo del 6%, cuando anteriormente eran establecidos como positivos para la prueba de la hemoglobina glucosilada [35]. En relación con esto, Mast y colaboradores (1998) [45] mencionan que los valores de ferritina menores de 12 µg/L poseen una sensibilidad del 25% y una especificidad del 98% para demostrar la ferropenia; sin embargo, cuando el valor límite para la ferritina se eleva a menores de 30 µg/L el valor predictivo positivo para indicar esta condición se incrementa a un 92%.

Por otra parte, los valores de sensibilidad y especificidad presentan la desventaja de que no proporcionan información relevante a la hora de tomar una decisión clínica ante un determinado resultado de una prueba de laboratorio; sin embargo, tienen la ventaja de que son propiedades intrínsecas a la prueba diagnóstica y definen su validez independientemente de cuál sea la prevalencia de la enfermedad en la población a la cual se aplica. Por el contrario, los valores predictivos, a pesar de ser de enorme utilidad a la hora de tomar decisiones clínicas y transmitir a los pacientes información sobre su diagnóstico, presentan la limitación de que dependen, en gran medida, de la frecuencia de la enfermedad a diagnosticar en la población objeto de estudio. De esta manera, cuando la prevalencia de la enfermedad es baja un resultado negativo permitirá descartar la enfermedad con mayor seguridad, lo que se observa como un valor predictivo negativo alto; mientras que un resultado positivo no permitirá confirmar el diagnóstico, lo que se refleja con un valor predictivo positivo bajo [46].

Al igual que en los estudios sobre tratamiento, todo resultado de un estudio de diagnóstico clínico o de laboratorio debe ser informado con su intervalo de confianza. El intervalo de confianza (IC) es el rango de valores dentro del cual se encuentra el valor verdadero, con un grado prefijado de certeza que habitualmente es del 95%. Así, se utiliza un «intervalo de confianza del 95%» o un «IC95%», que quiere decir que dentro de ese intervalo se encontrará el valor verdadero en un 95% de los casos. Cuanto más estrecho es el intervalo mayor confianza tendremos para utilizar el resultado [46].

La no inferioridad o concordancia en la comparación de las pruebas de laboratorio

El concepto de la no inferioridad se aplica más comúnmente en el campo de los ensayos clínicos, en especial en el campo de la terapéutica médica, para demostrar que un medicamento no es diferente a otro con mecanismo de acción y resultados clínicos similares, aunque puede serlo en cuanto a efectos adversos, modo de administración, costo, entre otros. Este mismo concepto puede ser extrapolado a las pruebas de laboratorio para determinar que una prueba nueva no es inferior a otra ya validada; es decir, para demostrar que los resultados promedio de la medición del mismo material biológico, obtenidos por dos pruebas diferentes, tienen una diferencia inferior a los límites predeterminados. En ese sentido, es permitido usar el término concordancia que se traduce en el grado en que dos o más observadores, métodos, técnicas u observaciones están de acuerdo sobre el mismo fenómeno observado [8].

La concordancia entre los métodos y sus mediciones se puede alterar por los siguientes elementos o fuentes de error: a) la variabilidad de los observadores, b) la variabilidad dada por el instrumento de medida y c) la variabilidad debida a medir en momentos diferentes en el tiempo, por lo que es esencial que cualquier nueva prueba se evalúe comparando su validez contra una prueba estándar de referencia (del inglés, *Gold Standard*) [8].

Durante la descripción de las características operativas de las nuevas pruebas se pueden introducir diferentes sesgos; uno de ellos es la inadecuada selección de la prueba estándar de referencia, ya que es frecuente que no se tenga una que sea perfecta, lo que lleva, en general, a sobreestimar las características operativas de las pruebas en comparación. Una medida utilizada para comparar dos pruebas es el **índice o valor kappa** de una prueba, el cual se calcula también a partir de la tabla binaria de 2x2 a la cual nos hemos referido anteriormente. Un valor de kappa de 0 indica que no existe concordancia entre las pruebas, descontado el factor azar, mientras que un valor de kappa de 1 indica una concordancia total de las pruebas [47]. El cálculo se realiza de la siguiente manera:

Concordancia esperada
=(a+b/n x a+c/n) + (c+d/n x b+d/n)

La máxima proporción de concordancia no debida al azar sería entonces:

Máxima proporción de concordancia
= 1-concordancia esperada

La concordancia descontando el azar sería:

Concordancia sin *azar*
= (a+d)/n-concordancia esperada

De esta manera, el valor kappa corresponde al cociente entre la proporción observada de concordancia descontando el azar y la máxima proporción de concordancia no debida al azar, expresado mediante la fórmula:

K = ((a+d)/n-concordancia esperada)/
(1-concordancia esperada)

Un valor de kappa se puede clasificar de la siguiente forma: de 0,2 a 0,4 indica una concordancia débil, 0,5 indica un nivel moderado de concordancia y un valor mayor de 0,8 indica una excelente concordancia [48]. De esta manera, la no inferioridad entre pruebas diagnósticas para una misma enfermedad puede ser calculada de dos formas. La primera es estableciendo límites predeterminados en una gráfica de sensibilidad (delta 1) y la tasa de falsos positivos (delta 2), si la prueba supera el límite para la sensibilidad se demuestra la superioridad y si supera el límite de tasa de falsos positivos no se acepta la no inferioridad. La segunda forma es calculando el índice kappa de cada prueba y si están en el mismo rango o nivel de concordancia se acepta la no inferioridad o equivalencia de las pruebas.

Veamos el siguiente ejemplo:

Para la determinación de la superioridad de la sensibilidad y la no inferioridad de la especificidad de las pruebas serológicas modificadas, respecto a las convencionales para el diagnóstico de *Trypanosoma cruzi* en la enfermedad de Chagas, se analizaron cinco pruebas: la inmunofluorescencia indirecta convencional y modificada, el ELISA convencional y modificada, y el Chagatek®. El índice kappa para la prueba Chagatek®, respecto a la prueba de ELISA modificada, fue de 0,87, y para esta última frente a la ELISA convencional fue de 0,96, lo que indica una excelente concordancia entre ambas pruebas y no inferioridad de las pruebas modificadas en relación con las convencionales [47].

Ilustraremos otro ejemplo:

Para el diagnóstico de la infección por *Trypanosoma cruzi* en fase crónica se utilizan pruebas serológicas, por lo que, en este caso, se quería evaluar y comparar la reproducibilidad de las pruebas de ELISA, inmunofluorescencia indirecta e inhibición de la hemaglutinación para el diagnóstico de la infección por *Trypanosoma cruzi* en mujeres embarazadas de una zona endémica para la enfermedad de Chagas en Colombia. Para esto se utilizaron muestras de suero y elución sanguínea, seleccionadas mediante muestreo de corte transversal [49].

Al comparar la primera lectura con la segunda de cada prueba practicada en las muestras de suero se encontró que la prueba de ELISA (punto de corte = 0,3), la inmunofluorescencia indirecta (punto de corte = 1/32) y la inhibición de la hemaglutinación (punto de corte = 1/16) presentaron índices kappa mayores de 0,8 (0,98, IC 95% = 0,93 a 1,00; 0,98, IC 95% = 0,92 a 1,00 y 0,88, IC 95% = 0,74-0,97, respectivamente), lo que indica que tienen una excelente concordancia entre las lecturas. Además, no se evidenciaron diferencias estadísticamente significativas entre las tres pruebas evaluadas (p>0,05), lo que comprueba la no inferioridad entre las pruebas.

En conclusión, las tres pruebas serológicas presentaron una reproducibilidad perfecta en suero, determinada mediante el índice kappa, por lo que cualquiera de ellas sería útil para establecer el diagnóstico de la infección por *Trypanosoma cruzi*. Por su simplicidad y su costo, la prueba ELISA se recomienda como prueba de elección para los programas de tamización de esta infección.

Retomando, la no inferioridad de una prueba (respecto a otra) también es posible demostrarla de forma descriptiva cuando el tamaño muestral es pequeño debido a una baja prevalencia de la enfermedad. En este caso, se requiere que los parámetros de sensibilidad y especificidad entre las pruebas sean similares, sin necesidad de realizar mayores cálculos matemáticos. Por ejemplo, para demostrar la no inferioridad para el diagnóstico de la tripanosomiasis africana de un prototipo de prueba diagnóstica rápida de inmunocromatografía (el SD BIOLINE HAT), que se basa en dos antígenos tripanosomales nativos, respecto al estándar de referencia (la prueba de hemaglutinación a una dilución 1/8), la cual requiere almacenamiento en frío y electricidad, se obtuvieron muestras de sangre de 14.818 sujetos de tres países de África central, endémicos para la enfermedad, de los cuales 149 fueron confirmados con tripanosomiasis africana por parasitología [50].

Al evaluar la sensibilidad y especificidad de la prueba SD BIOLINE HAT se encontraron valores del 89,26% (IC 95% = 83,27 a 93,28) y 94,58% (IC 95% = 94,20 a 94,94), respectivamente, mientras que para la prueba de hemaglutinación a una dilución 1/8 fueron del 89,26% (IC 95% = 83,27 a 93,28) y 98,88% (IC 95% = 98,70 a 99,04), respectivamente, lo que demostró una no inferioridad descriptiva entre estas dos pruebas [50].

Para una mayor comprensión sobre el índice o coeficiente kappa se sugiere remitirse a referencias bibliografías con una revisión más exhaustiva [51-53].

Conclusiones

La importancia de los parámetros para analizar los resultados de las pruebas de diagnóstico clínico y de laboratorio radica en su utilidad para diferenciar con certeza un paciente enfermo de uno sano. En consecuencia, una prueba diagnóstica ideal sería aquella que es capaz de detectar la mayor cantidad de pacientes con

la condición clínica estudiada, excluyendo, a la vez, a la mayor cantidad de pacientes sin ella.

Los resultados de una prueba diagnóstica nos ayudan a conocer la probabilidad de que un paciente determinado presente o no cierta enfermedad, de allí la relevancia de analizar estos resultados de acuerdo con la sensibilidad, la especificidad y los valores predictivos positivo y negativo. Además, el uso de la razón de verosimilitud brinda a los clínicos una mayor ayuda en el proceso diagnóstico, al hacer explícito el cambio entre la probabilidad pre y posprueba. En este sentido, la sensibilidad y especificidad de las pruebas diagnósticas también puede variar considerablemente.

Dado que los resultados de una prueba de diagnóstico pueden depender de las características propias de muchos pacientes, los estudios que reportan correlaciones diagnósticas deben, a su vez, proporcionar una descripción detallada de los diferentes subgrupos de pacientes. Del mismo modo, hay que considerar los efectos potenciales de la historia clínica, comorbilidades, protocolo de la prueba y extensión o grado de severidad de la enfermedad, para que el médico pueda interpretar los resultados de una prueba diagnóstica con mayor certeza [54].

Bibliografía

1. **Vizcaíno-Salazar G.** Sensibilidad y Especificidad. Medicina basada en la evidencia y análisis de diseños de investigación clínica. Maracaibo, Venezuela: EDILUZ; 2002: 57-71.
2. **Center for Evidence Basic Medicine (CEBM).** SpPin and SnNout. 2017. Disponible: http://www.cebm.net/sppin-and-snnout/. Consultado: jun 2017.
3. **Griner PF, Mayewski RJ, Mushlin AI, Greenland P.** Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981; 94: 557-592.
4. **Pita-Fernández S, Pértegas-Díaz S.** Pruebas diagnósticas: Sensibilidad y especificidad. Cad Aten Primaria 2003; 10: 120-124.
5. **Sierra-Arango F.** La sensibilidad y especificidad: entendiendo su origen y utilidad real. Rev Col Gastroenterol 2003; 18: 180-182.
6. **Rijneke RD, Ascoop CA, Talmon JL.** Clinical significance of upsloping ST segments in exercise electrocardiography. Circulation 1980; 61: 671-678.
7. **Ryder E, Gómez ME, Fernández V, Campos G, Morales LM, Valbuena H, et al.** Respuesta de la Glucosa/Insulina a una sobrecarga glucosada en sujetos con riesgo a diabetes tipo 2. Invest Clin 2001; 42: 255-268.
8. **Cortés-Reyes É, Rubio-Romero JA, Gaitán-Duarte H.** Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. Rev Colomb Obstet Ginecol 2010; 61: 247-255.
9. **Cerda J, Cifuentes L.** Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. Rev Chil Infectol 2012; 29: 138-141.
10. **Akobeng AK.** Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr 2007; 96: 644-647.
11. **Altman DG, Bland JM.** Diagnostic tests 3: receiver operating characteristic plots. BMJ 1994; 309: 188.
12. **Valenzuela D L, Cifuentes A L.** Validez de estudios de tests diagnósticos. Rev Med Chile 2008; 136: 401-404.
13. **López de Ullibarri G I, Pita-Fernández S.** Curvas ROC. Cad Aten Primaria 1998; 5 229-235.
14. **XLSTAT.** Curvas ROC. 2017. Addinsoft. Disponible: https://www.xlstat.com/es/soluciones/funciones/curvas-roc. Consultado: agosto 2017.
15. **Torres-Ortiz A.** Curvas ROC para Datos de Supervivencia. Aplicación a Datos Biomédicos. Tesis para optar al título de Master en Técnicas Estadísticas. Santiago de Compostela, España: Universidad de Santiago de Compostela; 2010.
16. **Gisbert JP, Ducons J, Gomollon F, Dominguez-Munoz JE, Borda F, Mino G, et al.** Validation of the 13c-urea breath test for the initial diagnosis of helicobacter pylori infection and to confirm eradication after treatment. Rev Esp Enferm Dig 2003; 95: 121-126, 115-120.
17. **Kroenke K, A Lucas C, Rosenberg M, Scherokman B, E Herbers J, A Wehrle P, et al.** Causes of persistent dizziness: A prospective study of 100 patients in ambulatory care. Ann Intern Med 1992; 117: 898-904.
18. **Salech F, Mery V, Larrondo F, Rada G.** Estudios que evalúan un test diagnóstico: interpretando sus resultados. Rev Med Chile 2008; 136: 1208-1208.
19. **Genders TS, Ferket BS, Hunink MG.** The Quantitative Science of Evaluating Imaging Evidence. JACC Cardiovasc Imaging 2017; 10: 264-275.
20. **Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M.** Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. Ann Intern Med 2010; 152: 167-177.
21. **Williams J W Jr, Simel D L.** Does this patient have sinusitis? Diagnosing acute sinusitis by history and physical examination. JAMA 1993; 270: 1242-1246.
22. **Williams J W Jr, Simel D L.** The rational clinical examination. Does this patient have ascites? How to divine fluid in the abdomen. JAMA 1992; 267: 2645-2648.
23. **Cook DJ, Simel DL.** The Rational Clinical Examination. Does this patient have abnormal central venous pressure? JAMA 1996; 275: 630-634.
24. **Turnbull JM.** The rational clinical examination. Is listening for abdominal bruits useful in the evaluation of hypertension? JAMA 1995; 274: 1299-1301.
25. **Udell JA, Wang CS, Tinmouth J, FitzGerald JM, Ayas NT, Simel DL, et al.** Does this patient with liver disease have cirrhosis? JAMA 2012; 307: 832-842.
26. **Berezin AE, Lisovaya OA.** C-reactive protein after stroke in arterial hypertension. Asian Cardiovasc Thorac Ann 2014; 22 551-557.
27. **Guyatt GH, Patterson C, Ali M, Singer J, Levine M, Turpie I, et al.** Diagnosis of iron-deficiency anemia in the elderly. Am J Med 1990; 88: 205-209.
28. **Cardona-Arias JA, Carrasquilla-Agudelo YE, Restrepo-Posada DC.** [Validity of three methods for inmuno-diagnostic of neurocysticercosis: systematic review of the literature with meta-analysis 1960-2014]. Rev Chilena Infectol 2017; 34: 33-44.

29. **Jin T, Fei B, Zhang Y, He X.** The Diagnostic Value of Polymerase Chain Reaction for Mycobacterium tuberculosis to Distinguish Intestinal Tuberculosis from Crohn's Disease: A Meta-analysis. Saudi J Gastroenterol 2017; 23: 3-10.

30. **Wu Y, Wang M, Zhu Y, Lin S.** Serum interleukin-6 in the diagnosis of bacterial infection in cirrhotic patients: A meta-analysis. Medicine (Baltimore) 2016; 95: e5127.

31. **Zacharias K, Ahmed A, Shah BN, Gurunathan S, Young G, Acosta D, et al.** Relative clinical and economic impact of exercise echocardiography vs. exercise electrocardiography, as first line investigation in patients without known coronary artery disease and new stable angina: a randomized prospective study. Eur Heart J Cardiovasc Imaging 2017; 18: 195-202.

32. **Chey WD, Wong BC.** American College of Gastroenterology guideline on the management of Helicobacter pylori infection. Am J Gastroenterol 2007; 102: 1808-1825.

33. **Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, et al.** Value of assessment of pretest probability of deep-vein thrombosis in clinical management. Lancet 1997; 350: 1795-1798.

34. **Gurusamy KS, Giljaca V, Takwoingi Y, Higgie D, Poropat G, Stimac D, et al.** Ultrasound versus liver function tests for diagnosis of common bile duct stones. Cochrane Database Syst Rev 2015: CD011548.

35. **Jarmul JA, Pignone M, Pletcher MJ.** Interpreting Hemoglobin A1C in Combination With Conventional Risk Factors for Prediction of Cardiovascular Risk. Circ Cardiovasc Qual Outcomes 2015; 8: 501-507.

36. **Best LM, Webb AC, Gurusamy KS, Cheng SF, Richards T.** Transcranial Doppler Ultrasound Detection of Microemboli as a Predictor of Cerebral Events in Patients with Symptomatic and Asymptomatic Carotid Disease: A Systematic Review and Meta-Analysis. Eur J Vasc Endovasc Surg 2016; 52: 565-580.

37. **Allen VB, Gurusamy KS, Takwoingi Y, Kalia A, Davidson BR.** Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer. Cochrane Database Syst Rev 2013; (11): CD009323.

38. **Cofre F, Delpiano L, Labrana Y, Reyes A, Sandoval A, Izquierdo G.** [TORCH syndrome: Rational approach of pre and post natal diagnosis and treatment. Recommendations of the Advisory Committee on Neonatal Infections Sociedad Chilena de Infectologia, 2016]. Rev Chilena Infectol 2016; 33: 191-216.

39. **Carpenter CR, Hussain AM, Ward MJ, Zipfel GJ, Fowler S, Pines JM, et al.** Spontaneous Subarachnoid Hemorrhage: A Systematic Review and Meta-analysis Describing the Diagnostic Accuracy of History, Physical Examination, Imaging, and Lumbar Puncture With an Exploration of Test Thresholds. Acad Emerg Med 2016; 23: 963-1003.

40. **Singh-Ospina N, Brito JP, Maraka S, Espinosa de Ycaza AE, Rodriguez-Gutierrez R, Gionfriddo MR, et al.** Diagnostic accuracy of ultrasound-guided fine needle aspiration biopsy for thyroid malignancy: systematic review and meta-analysis. Endocrine 2016; 53: 651-661.

41. **Aznar-Oroval E, Mancheño-Alvaro A, García-Lozano T, Sánchez-Yepes M.** Razón de verosimilitud y nomograma de Fagan: 2 instrumentos básicos para un uso racional de las pruebas del laboratorio clínico. Rev Calid Asist 2013; 28: 390-393.

42. **Fagan TJ.** Letter: Nomogram for Bayes theorem. N Engl J Med 1975; 293: 257.

43. **Ruiz-Canela Cáceres J, García Vera C.** Las estimaciones de validez de las pruebas diagnósticas varían según la prevalencia en los estudios. Evid Pediatr 2014; 10: 8.

44. **Best LM, Rawji V, Pereira SP, Davidson BR, Gurusamy KS.** Imaging modalities for characterising focal pancreatic lesions. Cochrane Database Syst Rev 2017; 4: CD010213.

45. **Mast AE, Blinder MA, Gronowski AM, Chumley C, Scott MG.** Clinical utility of the soluble transferrin receptor and comparison with serum ferritin in several populations. Clin Chem 1998; 44: 45-51.

46. **Bermejo Fraile B.** Validez de las pruebas diagnósticas. En: Epidemiología clínica aplicada a la toma de decisiones en medicina. Pamplona, España: Gobierno de Navarra, Departamento de Salud; 2001: 69-108.

47. **Vacca-Carvajal MA, Mercado-Reyes MM.** Determinación de las características operativas de las pruebas serológicas con cepas colombianas de Trypanosoma cruzi utilizadas para el diagnóstico de enfermedad de Chagas. Tesis para optar al título de Magister en Epidemiología Clínica. Bogotá, Colombia: Pontificia Universidad Javeriana; 2005.

48. **Diamond GA.** Clinical epistemology of sensitivity and specificity. J Clin Epidemiol 1992; 45: 9-13.

49. **Castellanos YZ, Cucunubá ZM, Flórez AC, Orozco-Vargas LC.** Reproducibilidad de pruebas serológicas para el diagnóstico de infección por Trypanosoma cruzi en mujeres embarazadas de una zona endémica de Santander, Colombia. Biomédica 2014; 34: 198-206.

50. **Bisser S, Lumbala C, Nguertoum E, Kande V, Flevaud L, Vatunga G, et al.** Sensitivity and Specificity of a Prototype Rapid Diagnostic Test for the Detection of Trypanosoma brucei gambiense Infection: A Multi-centric Prospective Study. PLoS Negl Trop Dis 2016; 10: e0004608.

51. **Gwet KL.** The Kappa Coefficient: A Review. En: Handbook of Inter-Rater Reliability (ed 3a). Maryland, Estados Unidos: Advanced Analytics, LLC; 2012: 15-46.

52. **Carrasco JL, Jover L.** Métodos estadísticos para evaluar la concordancia. Med Clin 2004; 122: 28-34.

53. **Chmura Kraemer H, Periyakoil VS, Noda A.** Kappa coefficients in medical research. Stat Med 2002; 21: 2109-2129.

54. **Hlatky MA, Pryor DB, Harrell FE, Jr., Califf RM, Mark DB, Rosati RA**. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984; 77: 64-71.

Abstract: *The selection of a test to request a patient as well as their interpretation is a daily scenario in which the physician must deal and apply his critical judgment based on the reported evidences. It is common that when talking about a clinical or laboratory diagnostic test, parameters such as sensitivity, specificity and positive and negative predictive values are described. These parameters reflect the characteristics of a diagnostic test and serve to decide when should be used (sensitivity and specificity of a test) or what is the meaning of a test result in a particular patient. When it is necessary to compare these parameters in different tests and opting for the most useful for the diagnosis of a particular disease, it is essential that the physician knows and learns how these measures are obtained and their interpretation to decide the most appropriate behavior for the patient. The objective of the present review is to provide basic and simple statistical concepts for the understanding and application of clinical and laboratory diagnostic tests.*

Key words: *Sensitivity and specificity, predictive value of tests, ROC curve, likelihood functions, pre- and post-test probability, prevalence.*