PRUEBAS DIAGNÓSTICAS

ÍNDICE

PRUEBAS DIAGNÓSTICAS

CONCEPTOS GENERALES

El diagnóstico puede considerarse como el más importante resultado de la práctica médica, la clave que conduce al tratamiento y al pronóstico. Resulta, asimismo, un problema complejo en ese ejercicio, ya que en ocasiones resulta inalcanzable y a veces, paradójicamente, innecesario.

El diccionario Webster lo define como: 1) el acto o proceso de determinar la naturaleza de la condición mórbida mediante el examen; 2) un cuidadoso examen de los hechos para determinar la naturaleza de algo y 3) la decisión u opinión resultante de tal examen o investigación. Por su parte Kassirer, uno de los autores que más ha publicado sobre este tema, señala que el diagnóstico es una hipótesis acerca de la naturaleza de la enfermedad de un paciente que se deriva de observaciones a través del uso de la inferencia[1]. Sobre su significado, algunos autores concluyen que el diagnóstico es un resultado de alta significación para el médico, pero mucho más lo es para el paciente[1,2]. Según Gaarder[2], para el médico el diagnóstico es un objetivo ideal y elusivo (dispuesto a recordarnos nuestras limitaciones), mientras que para el paciente significa eliminar la incertidumbre de saber que algo anda mal en él y no saber qué es.

Por otro lado, la mayoría de los autores reconocen que la presencia de una enfermedad en un individuo a menudo no puede determinarse con certeza[1,3]. Kassirer es aún más categórico en este aspecto y refiere que la certeza absoluta en diagnóstico es inalcanzable, independientemente de cuánta información se obtenga, cuántas observaciones se hagan, o cuántas pruebas diagnósticas se realicen en el desempeño médico. Agrega, además, que el objetivo del médico no es alcanzar la certeza sino reducir el nivel de incertidumbre lo suficiente como para tomar la decisión terapéutica[1].

De cómo transcurre el proceso que da lugar al diagnóstico se ocupa también la literatura, varios autores coinciden en que éste requiere de dos etapas diferenciales. En la primera se establece una presunción, sospecha o hipótesis de existencia de la enfermedad[4-7]. La segunda se dirige al seguimiento de la suposición clínica y a verificar si la hipótesis corresponde a la verdad[4,5].

En este proceso, de compleja estructura, existe un gran número de fuentes de incertidumbre que transitan por una amplia gama de cuestiones como son: que el conjunto de síntomas y signos en un paciente puede ser compatible con más de una enfermedad, que existen variaciones biológicas a veces importantes entre un enfermo y otro, que los instrumentos suelen ser imprecisos, y que los pacientes son inexactos para recordar sucesos pasados.

Durante todo el proceso que conduce al diagnóstico, el médico se vale de distintas fuentes de información. Se destacan entre ellas la anamnesis del paciente, el examen físico, la información epidemiológica y los resultados de las llamadas pruebas diagnósticas. Es sobre estas últimas que se centrará la atención en este módulo.

Se llamará ***prueba diagnóstica*** (PD) a cualquier proceso, más o menos complejo, que pretenda determinar en un paciente la presencia de cierta condición, supuestamente patológica, no susceptible de ser observada directamente (con alguno de los cinco sentidos elementales). Es decir, que no se suelen considerar como pruebas diagnósticas a los sentidos cuando evalúan la presencia de algún signo patológico. Si se observa un aumento de volumen en una extremidad por ejemplo, no se considera esa observación como el "diagnóstico de un aumento de volumen"; pero si con esa observación se deduce que el paciente tiene un "melanoma maligno", entonces la observación si está

actuando como PD, ya que el "melanoma maligno" no puede observarse directamente. La definición se refiere a "condición" y no enfermedad o entidad gnosológica, ya que no siempre se utiliza una PD para identificar una enfermedad, sino que ésta también puede utilizarse para diagnosticar síndromes o procesos patológicos.

Mucho se ha escrito en torno a las pruebas diagnósticas y a su eficacia real como elementos contribuyentes a la correcta clasificación diagnóstica de un paciente en estudio. Silva[8] y Begg y Greenes[9], por ejemplo, afirman que el uso de pruebas diagnósticas para la detección y evaluación de varias enfermedades en la práctica médica, ha crecido notablemente en años recientes, y tiende a incrementarse exponencialmente. El desarrollo tecnológico de los últimos decenios ha permitido incorporar a la práctica clínica médica novedosos y sofisticados medios diagnósticos que, sin duda, constituyen adelantos en el perfeccionamiento del trabajo médico. Lamentablemente, estos adelantos en los medios diagnósticos se han acompañado también de una tendencia a su uso indiscriminado. Ante una hipótesis diagnóstica y un conjunto (a veces numeroso) de pruebas que ayudan a corroborarla, el médico no siempre se propone hacer de ellos un uso racional.

La necesidad de herramientas cuantitativas que contribuyan a dirigir con racionalidad las indicaciones es indispensable. Se trata, en particular, de obtener índices o medidas de eficacia de cada medio diagnóstico que sirvan de pauta orientadora para su selección en el momento necesario. Se parte de la premisa de que, en cada momento, el médico deberá hacer un uso racional de los distintos instrumentos y procedimientos que le son útiles para llegar al diagnóstico.

El módulo de Pruebas Diagnósticas de Epidat 3.1 permite hacer un uso eficiente de las herramientas cuantitativas principales existentes para evaluar la eficacia de las pruebas diagnósticas y contribuir a su uso racional.

INDICADORES ESTADÍSTICOS BÁSICOS PARA EVALUAR EL DESEMPEÑO DE UN PROCEDIMIENTO DIAGNÓSTICO

La evaluación del desempeño de una prueba diagnóstica comienza por la cuantificación (estimación, más bien) de la magnitud de los errores que pueden cometerse o, su inverso, la magnitud de los aciertos que se cometen al intentar "adivinar" un diagnóstico a partir de los resultados que brinde dicho procedimiento.

Sensibilidad y especificidad

En 1947, Yerushalmy[♠] introduce los términos de sensibilidad y especificidad como indicadores estadísticos que evalúan el grado de eficacia inherente a una prueba diagnóstica (citado en 4,10).

La **sensibilidad** y la **especificidad** son las medidas tradicionales y básicas del valor diagnóstico de una prueba. Miden la discriminación diagnóstica de una prueba en relación a un criterio de referencia, que se considera la verdad.

Estos indicadores en principio permiten comparar directamente el eficacia de una prueba con el de otras y esperar resultados similares cuando son aplicadas en diferentes países, regiones o ámbitos.

[♠] La referencia al artículo original de Yerushalmy es la siguiente:
Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Pub Health Rep* 1947; 62: 1432-49.

La **sensibilidad** (S) indica la capacidad de la prueba para detectar a un sujeto enfermo, es decir, expresa cuan "sensible" es la prueba a la presencia de la enfermedad[4,6,8,10-18]. Para cuantificar su expresión se utilizan términos probabilísticos: si la enfermedad está presente, ¿cuál es la probabilidad de que el resultado sea positivo?

La respuesta es una expresión en términos de probabilidad condicional:

S = P (T+/Enf)

o sea, la sensibilidad es la probabilidad de que la prueba identifique como enfermo a aquél que efectivamente lo está.

La **especificidad** (E) indica la capacidad que tiene la prueba de identificar como sanos (no enfermos) a los que efectivamente lo son[4,6,8,10-18]. Se define entonces también como la probabilidad condicional:

E = P (T-/no Enf)

es decir, la especificidad es la probabilidad de que la prueba identifique como no enfermo a aquél que efectivamente no lo está.

T+ y **T-** indican, respectivamente, un resultado positivo o negativo de la prueba o test diagnóstico.

Estimación de S y E

Para ilustrar el significado de estos conceptos a través de sus estimaciones, supóngase que se tienen N sujetos de los que se conoce su estatus verdadero (enfermo o no) y se les ha practicado el test o prueba que se está evaluando y cuyo resultado puede ser inequívocamente positivo o negativo.

Estas características pueden entonces estimarse fácilmente a partir de una tabla de 2x2 como se muestra a continuación:

Tabla 1. Resultados de la prueba y la existencia de la enfermedad.

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba diagnóstica	**Positivos**	a	b	a+b
	Negativos	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Donde:

a = número de pacientes con la enfermedad diagnosticados como "positivos" por la prueba.

b = número de pacientes sin la enfermedad diagnosticados como "positivos" por la prueba.

c = número de pacientes con la enfermedad diagnosticados como "negativos" por la prueba.

d = número de pacientes sin la enfermedad diagnosticados como "negativos" por la prueba[♠].

Puede apreciarse que cada celda de la tabla refleja una característica que también suele calificarse de la manera siguiente:

a = Verdaderos positivos (VP)

b = Falsos positivos (FP)

[♠] Por la frecuencia con que se mencionará esta tabla se le identificará como "tabla básica 2x2".

c = Falsos negativos (FN)

d = Verdaderos negativos (VN)

Con estos términos, la tabla puede expresarse así:

Tabla 2. Resultados de la prueba y la existencia de la enfermedad

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba diagnóstica	**Positivos**	VP	FP	VP+FP
	Negativos	FN	VN	FN+VN
	Total	VP+FN	FP+VN	N = (VP+FP+FN+VN)

Por tanto, los estimadores de las probabilidades descritas son, naturalmente, los siguientes:

$$Sensibilidad = \frac{Verdaderos\ positivos}{Total\ de\ enfermos} = \frac{VP}{VP + FN} = \frac{a}{a + c}$$

$$Especificidad = \frac{Verdaderos\ negativos}{Total\ de\ no\ enfermos} = \frac{VN}{VN + FP} = \frac{d}{b + d}$$

Valores predictivos

A pesar de que la S y la E se consideran las características operacionales fundamentales de una prueba diagnóstica, en la práctica su capacidad de cuantificación de la incertidumbre médica es limitada. El médico necesita más bien evaluar la medida en que sus resultados modifican realmente el grado de conocimiento que se tenía sobre el estado del paciente. Concretamente, le interesa conocer la probabilidad de que un individuo para el que se haya obtenido un resultado positivo, sea efectivamente un enfermo; y lo contrario, conocer la probabilidad de que un individuo con un resultado negativo esté efectivamente libre de la enfermedad. Las medidas o indicadores que responden a estas interrogantes se conocen como **valores predictivos.**

El **valor predictivo de una prueba positiva** equivale a la probabilidad condicional de que los individuos con una prueba positiva tengan realmente la enfermedad[4,6,8,10-18]:

VP(+) = P(Enf/T+)

El **valor predictivo de una prueba negativa** es la probabilidad condicional de que los individuos con una prueba negativa no tengan realmente la enfermedad[4,6,8,10-18]:

VP(-) = P(No Enf/T-)

Mediante la tabla de 2x2 que se introdujo antes se puede ilustrar también cómo se estiman los valores predictivos (suponiendo que esta tabla se conforme seleccionando una muestra al azar de tamaño N de la población, y luego se clasifiquen los sujetos de la muestra en los cuatro grupos posibles según la prueba diagnóstica y el criterio de verdad):

$$Valor\ predictivo\ positivo = \frac{Verdaderos\ positivos}{Total\ de\ positivos} = \frac{VP}{VP + FP} = \frac{a}{a + b}$$

$$Valor\ predictivo\ negativo = \frac{Verdaderos\ negativos}{Total\ de\ negativos} = \frac{VN}{VN + FN} = \frac{d}{c + d}$$

Diseños a utilizar para la estimación de los indicadores. Ventajas y desventajas

Hasta aquí se ha considerado la manera de calcular, o estimar puntualmente, los indicadores básicos para evaluar la eficacia de una PD. Se ha dicho poco sobre cómo obtener los datos, o lo que es lo mismo, cómo diseñar el estudio para obtener los datos de los pacientes. De hecho se ha dado por sentado que "de alguna manera" se tiene una muestra de tamaño N de cierta población a la cual se le ha aplicado el test a prueba y el criterio de verdad para hacer las estimaciones. Sin embargo, la organización de los sujetos que conformarán la muestra puede hacerse de varias maneras, las que se comentarán a continuación.

La vía canónica consiste en seleccionar una muestra de N_1 enfermos y otra de N_2 sujetos no enfermos diagnosticados por la prueba de referencia, y aplicar la nueva prueba a los $N = N_1 + N_2$ individuos de suerte que pueda conformarse una clasificación cruzada tal como se expone en una tabla de 2x2.

La tabla queda conformada como se expone a continuación:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba diagnóstica	**Positivos**	a	b	a+b
	Negativos	c	d	c+d
	Total	N_1	N_2	$N=N_1+N_2$

Los estimadores obvios de S y E vienen dados por las proporciones siguientes:

$$S = \frac{a}{N_1} \text{ y } E = \frac{d}{N_2}$$

Varias observaciones son pertinentes en este punto. En primer lugar, hay que enfatizar que el criterio de clasificación de los sujetos como enfermos o no, debe ser independiente de la prueba diagnóstica en estudio; el criterio (o la prueba) tomado como criterio de verdad se supone que tiene sensibilidad y especificidad máximos (ambos del 100%). Por otra parte, debe notarse que el diseño planteado no consiente la estimación de la prevalencia de la enfermedad P(E) a través de la razón N_1/N puesto que N_1 y N_2 son números elegidos por el evaluador, con frecuencia iguales entre sí. Consecuentemente, tampoco es posible en este caso estimar los valores predictivos, a menos que se conozca la prevalencia de la enfermedad en cuestión, o sea, que se cuente con una estimación independiente del parámetro P(E) y pueda entonces aplicarse el Teorema de Bayes (como se verá más adelante).

Si se quieren conocer los valores predictivos de una prueba y no se conoce la prevalencia en el contexto donde piensa utilizarse, entonces es necesario buscar otro diseño. Uno expresamente ideado para estimar valores predictivos consiste en seleccionar N(+) pacientes con una prueba positiva y N(-) con una prueba negativa y aplicarles a los N=N(+)+N(-) pacientes la prueba de referencia o criterio de verdad.

La tabla queda conformada de la siguiente manera:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba diagnóstica	**Positivos**	a	b	N(+)
	Negativos	c	d	N(-)
	Total	a+c	b+d	N=N(+)+N(-)

Los estimadores de VP(+) y VP(-) vienen dados por las proporciones siguientes:

$$VP(+) = \frac{a}{N(+)} \text{ y } VP(-) = \frac{d}{N(-)}$$

Pero este diseño no permite estimar S y E.

Un diseño cómodo para estimar los cuatro indicadores consiste en obtener una muestra simple aleatoria de N individuos y aplicar a cada uno de ellos el criterio de verdad o prueba de referencia con que se cuenta y la prueba diagnóstica que se evalúa. En tal caso todos los totales marginales son aleatorios, por lo que la estimación de los cuatro índices es aceptable.

La tabla 2x2 quedaría conformada como se expone a continuación:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba diagnóstica	**Positivos**	a	b	a+b
	Negativos	c	d	c+d
	Total	a+c	b+d	N

La estimación de los parámetros se expresa en las fórmulas siguientes:

$$S = \frac{a}{a+c} \text{ y } E = \frac{d}{b+d}$$

$$VP(+) = \frac{a}{a+b} \text{ y } VP(-) = \frac{d}{c+d}$$

En esta situación, la prevalencia de la enfermedad puede ser estimada por los resultados de la prueba de referencia:

$$P(E) = \frac{a+c}{N}$$

y los parámetros estadísticos apropiados pueden ser computados.

Tal estrategia, sin embargo, es raramente factible. Efectivamente, téngase en cuenta que lo que ocurre generalmente es que el criterio estándar de comparación es un método sofisticado, invasivo o económicamente prohibitivo, en tanto que la prueba que se evalúa se propone precisamente como un sucedáneo ventajoso de ser aplicado bajo condiciones de apremio de tiempo o de recursos.

Además, este último diseño tiene la desventaja de que si la muestra no es lo suficientemente grande, alguno de los parámetros puede quedar mal representado. Por ejemplo, si la prueba es poco sensible y la enfermedad infrecuente la sensibilidad puede quedar mal estimada por escaso tamaño de muestra.

El diseño, a juicio nuestro, que debe ser más utilizado, teniendo en cuenta costo, tiempo, recursos, riesgos en términos económicos y de daños para el paciente, es el primero explicado en este acápite para la estimación de S y E, que consiste en seleccionar una muestra de enfermos y otra de no enfermos diagnosticados por la prueba de referencia. Siempre que sea conocida *a priori* la prevalencia de la enfermedad pueden estimarse los valores predictivos positivo y negativo según se verá más adelante y, de esta forma, se estimarían los cuatro parámetros propuestos por Yerushalmy hace más de cuatro décadas, que sintetizan el valor intrínseco de una prueba y constituyen una vía, por ejemplo, para comparar dos o más pruebas que compiten entre sí.

Probabilidades pre y post prueba y Teorema de Bayes

El concepto epidemiológico puro, indica que prevalencia o tasa de prevalencia (términos en principio equivalentes) es la proporción de la población de individuos que en un lapso dado y una región determinada se consideran enfermos.

Pero desde la perspectiva de la actividad clínica, en principio más individualizada, la prevalencia de una enfermedad corresponde a la estimación de la probabilidad de padecer la enfermedad antes de realizar la prueba. En el ámbito clínico, la "prevalencia" se conoce como probabilidad anterior (*a priori*) a la prueba, es decir, la prevalencia de la enfermedad en una población de pacientes con las mismas características que el que se está evaluando. El valor predictivo significa lo mismo que la probabilidad de que la enfermedad esté presente (o ausente) después de obtener los resultados de la prueba. Por esta razón, el valor predictivo positivo puede considerarse clínicamente como la probabilidad de tener la enfermedad una vez que se tiene un resultado positivo (o negativo) o probabilidad *a posteriori*.

Conociendo la S y la E del test, los VP se pueden obtener (a partir del Teorema de Bayes) para las distintas circunstancias de la práctica médica.

Teorema de Bayes

Como se dijo antes, generalmente, al realizar la validación de un medio diagnóstico se reúne a un grupo de pacientes enfermos y a otro de no enfermos (clasificados según un criterio de verdad conocido); pero en este caso los valores predictivos no pueden ser estimados directamente de los datos por las razones ya planteadas.

Sin embargo, si se conoce la prevalencia, o probabilidad *a priori*, de la enfermedad en el contexto donde se aplicará la prueba, y se tienen la sensibilidad (S) y especificidad (E) de dicha prueba diagnóstica, se puede estimar los valores predictivos para ese contexto aplicando ciertas expresiones o fórmulas matemáticas que se derivan de la aplicación del conocido Teorema de Bayes.

Este teorema fue propuesto y demostrado en el siglo XVIII por el británico Tomas Bayes, quien pereció antes de que fuera publicado, y cuyos trabajos fueron reportados póstumamente por un amigo[18]. En sus trabajos originales Bayes desarrolló las fórmulas con el uso de las probabilidades condicionales y simbología probabilística.

Véase el planteamiento general de Bayes.

Sea A un suceso dado y, sean H_1, H_2, ..., H_k sucesos excluyentes que no contienen a A, pero que de alguna forma se suponen relacionados con él. Se supone que se conocen las probabilidades (*a priori*) de estos sucesos, $P(H_i)$, i = 1, 2, ..., k y las probabilidades condicionadas $P(A/H_i)$[19]. Entonces, las probabilidades a posteriori se pueden expresar como:

$$P(H_j \mid A) = \frac{P(A \mid H_j)P(H_j)}{\sum_{i=1}^{k} P(A \mid H_i)P(H_i)}$$

En el contexto de la evaluación de medios diagnósticos, se tienen los elementos siguientes:

- La prevalencia de la enfermedad que se denota como **P(Enf)** o P
- La prevalencia de no enfermos que se denota como **P(No Enf)** o Q=1-P
- La sensibilidad conocida, que puede denotarse como **P(T+/Enf)**, y que representa la probabilidad de que el test resulte positivo, dado que el paciente tiene la enfermedad.
- La especificidad conocida de un test, que puede denotarse como **P(T-/No Enf)**, o la probabilidad de que el test resulte negativo dado que el paciente "no" tiene la enfermedad.

Se tiene, entonces, dos sucesos excluyentes (enfermo y no enfermo). Si se toma el suceso "test positivo" entonces, a partir del Teorema de Bayes, el valor predictivo de una prueba positiva se escribe como **P(Enf/T+),** que representa la probabilidad de estar enfermo, dado que el paciente tiene un resultado positivo en la prueba diagnóstica[18].

Igualmente, a partir del suceso "test negativo", el valor predictivo de una prueba negativa se escribe como **P(NoEnf/T-)**, y representa la probabilidad de no estar enfermo, dado que el paciente tiene un resultado negativo en la prueba diagnóstica[18].

Entonces, aplicando el Teorema de Bayes:

$$P(Enf \mid T+) = \frac{P(T+ \mid Enf) \times P(Enf)}{P(T+ \mid Enf) \times P(Enf) + P(T+ \mid No\ Enf) \times P(No\ Enf)}$$

$$VP(+) = \frac{S \times P(Enf)}{S \times P(Enf) + (1 - E) \times [1 - P(Enf)]}$$

$$P(No\ Enf \mid T-) = \frac{P(T- \mid No\ Enf) \times P(No\ Enf)}{P(T- \mid No\ Enf) \times P(No\ Enf) + P(T- \mid Enf) \times P(Enf)}$$

$$VP(-) = \frac{E \times [1 - P(Enf)]}{E \times [1 - P(Enf)] + (1 - S) \times P(Enf)}$$

Con estas fórmulas pueden estimarse los valores predictivos en cualquier contexto poblacional (del ámbito epidemiológico tradicional) o específico (de la clínica).

Estimación por intervalos de confianza de la sensibilidad y la especificidad

Hasta el momento se ha considerado virtualmente el significado de S y E desde una óptica descriptiva: la S y la E calculadas (estimadas puntualmente) a partir de una tabla 2x2. No obstante, lo que se desea es estimar una propiedad genérica del test, y la tabla de 2x2 de donde se obtienen las estimaciones no es más que la expresión organizada de los datos de una muestra de una población que, generalmente, es de las clasificadas como "infinitas".

La sensibilidad y la especificidad son proporciones y en el contexto de la evaluación de medios diagnósticos también son válidos los conceptos de población y muestra y las generalidades del uso de la inferencia estadística. Se trata de utilizar los métodos inferenciales que se usan comúnmente en el área de análisis de datos cualitativos, ahora para la inferencia sobre proporciones.

Un intervalo de confianza (IC) conservador para una proporción p puede obtenerse empleando la aproximación a la Normal, que es el método empleado por Epidat 3.1 para el cálculo de los IC. Detalles sobre el procedimiento estadístico y las fórmulas empleadas pueden encontrarse en el texto de Fleiss[20].

Esta es la interpretación elemental de un IC con nivel de confianza 1-α: de cada 100 muestras que se obtengan de la misma población y se emplee el mismo procedimiento para obtener el IC, se espera que en (1-α)100% de ellos se encontrará realmente el parámetro (la S) y en α100% de ellos no.

Por ejemplo si, como es muy frecuente, α=0,05 entonces, de cada 100 muestras de la misma población (en este caso, de enfermos si es S lo que se estima, o de no enfermos si lo es E) donde se emplee el mismo método de cálculo, en 95 se encontrará el parámetro y en 5 no.

Otros indicadores del desempeño de un test

Si bien es cierto que la S, la E y los VP pueden considerarse los índices fundamentales en la evaluación de la bondad de un test, ellos no son los únicos índices conocidos. En primer lugar, tienen la desventaja de ser cuatro, o sea, en general resulta engorroso tener que caracterizar una prueba diagnóstica, muchas veces sencilla intrínsecamente, con cuatro indicadores distintos. En segundo lugar, a pesar de que permiten un conocimiento casi completo de la capacidad diagnóstica de un test y de su desempeño en la práctica, no abordan todas las aristas posibles en esa evaluación.

La literatura recoge varias proposiciones de indicadores posibles, algunos únicos, para evaluar el desempeño de un test diagnóstico. Se describen aquí tres de estos indicadores, que se pueden calcular con ayuda de Epidat 3.1.

Índice de validez o proporción correcta de aciertos (IV). Se define como la proporción de individuos clasificados correctamente. En términos de la tabla 2x2 básica, el índice de validez responde a la siguiente fórmula:

$$IV = (a+d)/N$$

Feinstein[10] demuestra cómo este índice depende, no solamente de la sensibilidad y la especificidad, sino también de la prevalencia de la enfermedad. En efecto, si se escriben los términos de la tabla básica como:

$a = S \times n_1$, donde $n_1 = a+c$

$d = E \times n_2$, donde $n_2 = b+d$

entonces:

P = Prevalencia de la enfermedad = n_1/N, y

Q = (1-P) = n_2/N

de modo que:

$$IV = (S\times n_1+E\times n_2)/N = S\times(n_1/N)+E\times(n_2/N) = S\times P+E\times(1-P) = S\times P+E-E\times P = P\times (S-E)+E$$

que representa la ecuación de una línea recta con intercepto en E y pendiente igual a la diferencia entre S y E. A medida que la prevalencia cambia, el IV se ve afectado (linealmente), independientemente de la S y la E, aunque mientras mayor es la diferencia S-E, también es más fuerte la dependencia de P. Si la diferencia es nula, el índice de validez será igual a la especificidad de la prueba.

El índice de validez rara vez es usado actualmente por su "falta de validez", a pesar de que es realmente atractivo por su sencillez.

Índice de Youden o versión 2 de la probabilidad corregida de detectar enfermedad (IJ). Una medida conjunta de eficiencia de un medio diagnóstico fue propuesta por W.J. Youden en 1950. Su estructura algebraica es la siguiente:

$$IJ = S+E-1 = S-(1-E)$$

Simplemente refleja la diferencia entre la tasa de verdaderos positivos y la de falsos positivos. Un buen test debe tener alta esta diferencia. Teóricamente es igual a 1 sólo cuando la prueba diagnóstica es perfecta, o sea, cuando S + E = 2, de modo que también puede decirse que cuánto más cercano a 1, mejor es la prueba diagnóstica que se está evaluando.

El IJ tiene la ventaja de no estar afectado por la selección de la prevalencia, y es preferido por la combinación de los sencillos valores de la sensibilidad y la especificidad[10,18]. Sin embargo, tiene la desventaja de que, al resultar de la combinación de los valores de S y E, se pierde la idea de si la prueba diagnóstica es buena en sensibilidad o especificidad. Feinstein[10] fundamenta esta afirmación mediante un ejemplo: si el índice de Youden tiene un valor de 0,55, puede ser que la sensibilidad sea de 0,95 y la especificidad de 0,60, o viceversa.

La razón de verosimilitud (RV). Feinstein[18] califica a la razón de verosimilitud[♠] como un indicador "reciente y popular" del desempeño de un test diagnóstico. La definición y, a la vez, la expresión matemática que parece más conocida es la siguiente:

$$RV+ = \frac{Sensibilidad}{1 - Especificidad}$$

Si se recuerdan las definiciones básicas de S y E se tiene:

$$RV+ = \frac{P(T+ \mid Enf)}{P(T+ \mid No\ Enf)}$$

que responde a la pregunta: ¿Cuántas veces más probable es que el test sea positivo en los enfermos que en los no enfermos?, una noción sugestiva, similar al concepto de riesgo relativo tan utilizado en la Epidemiología moderna.

[♠] *Likelihood ratio* es su nombre en inglés y las siglas "LR" la identifican en casi toda la literatura en lengua inglesa.

De este concepto es evidente que se desprende el complemento: la respuesta a la pregunta ¿cuántas veces más probable es que el test sea negativo en los enfermos que en los no enfermos? La respuesta es el cociente:

$$RV- = \frac{P(T- \mid Enf)}{P(T- \mid No\ Enf)}$$

llamada razón de verosimilitud para resultados negativos (lo que explica el signo negativo en la expresión anterior).

Si no se responde a esta pregunta no se tendrá una idea completa de la eficacia del test porque puede que un resultado positivo sea más probable en los enfermos que en los no enfermos (RV+ alto), pero con una especificidad menor de 0,5 la probabilidad de resultados negativos también será mucho mayor en los no enfermos que en los enfermos.

Véase el siguiente ejemplo hipotético:

	Enfermos	**No enfermos**
Test +	48	25
Test -	2	25
Total	50	50

RV+ = (48/50)/(25/50) = 1,92

RV- = (2/50)/(25/50) = 0,08

La probabilidad de un resultado positivo es, aproximadamente, dos veces mayor en los enfermos que en los no enfermos, pero la probabilidad de un resultado negativo es 12 veces mayor en los no enfermos que en los enfermos (1/0,08=12,5). Este test tiene una S alta pero una E muy baja, y esto se refleja en que la RV+ es sustancialmente mayor que la RV-, lo que le confiere mayor valor para detectar no enfermos que para detectar enfermos (los falsos negativos son improbables).

Un buen test debe tener una RV- cercana a 0 y una RV+ alta (no es posible especificar un límite superior para la RV+).

En resumen, la razón de verosimilitud combina la información que proviene de la sensibilidad y la especificidad y es definida como la razón entre la probabilidad de un resultado de una prueba en sujetos enfermos y la probabilidad del mismo resultado en sujetos no enfermos[12].

Puede ser, incluso, que una prueba tenga más de dos posibles resultados. Entonces, la razón de verosimilitud separada puede ser calculada para cada resultado Tx:

$$RVx = \frac{P(Tx \mid Enf)}{P(Tx \mid No\ Enf)}$$

expresión que le confiere a la RV un nivel de generalidad mucho mayor y da lugar a la llamada RV de un resultado específico, que permite conocer rápidamente si determinado resultado permite distinguir enfermos de no enfermos.

Como ya se ha visto, la RV es independiente de la prevalencia de la enfermedad, lo que constituye su principal virtud.

Ejemplo

Lo planteado anteriormente brinda la base teórica para trabajar. Véase lo que brinda Epidat 3.1 mediante un ejemplo tomado de Luck y col[21].

Supóngase que se tiene una situación donde se quiere estimar la sensibilidad y la especificidad de cierto cuestionario para diagnosticar la presencia de un desorden alimentario en adolescentes. El cuestionario tiene 5 preguntas que se responden con Sí o No y se considera positivo si la respuesta es Sí en al menos dos de las preguntas. Se toman 341 mujeres entre 15 y 25 años que acuden a una consulta de Psiquiatría durante 1 año. A todas cierto investigador les aplica el cuestionario en cuestión. Un año después, en el Servicio se tienen elementos suficientes (sin conocer el resultado de la aplicación del cuestionario) para clasificarlas a todas como "enfermas con un desorden alimentario de cualquier tipo" o "no enfermas de esa dolencia" y se obtiene la siguiente tabla de 2x2:

		Criterio de verdad (después del año de seguimiento)		
		Enfermas	**No enfermas**	**Total**
Prueba diagnóstica	**Positivas**	11	34	45
	Negativas	2	294	296
	Total	13	328	341

Para resolver el problema con Epidat 3.1 (módulo Pruebas diagnósticas, submódulo Pruebas simples), se escoge la opción "datos tabulados" y se introduce en la tabla que aparece en la pantalla los datos tal y como aparecen en la tabla anterior. Note que no es necesario introducir los totales. Se especifica el nivel del intervalo de confianza (95% es el preestablecido por ser el de mayor uso) y se presiona la tecla "calcular". Aparece la siguiente lista de resultados:

```
Pruebas diagnósticas simples

       Nivel de confianza:         95,0%

                       Prueba de referencia
  Prueba diagnóstica    Enfermos       Sanos       Total
--------------------   ---------   ---------   ---------
            Positivo          11          34          45
            Negativo           2         294         296
--------------------   ---------   ---------   ---------
               Total          13         328         341


                                 Valor            IC (95%)
-------------------------    ----------  ------------------------
         Sensibilidad (%)         84,62       61,16       100,00
       Especificidad (%)          89,63       86,18        93,09
    Indice de Validez (%)         89,44       86,03        92,85
   Valor predictivo + (%)         24,44       10,78        38,11
   Valor predictivo - (%)         99,32       98,22       100,00
          Prevalencia (%)          3,81        1,63         5,99

         Indice de Youden          0,74        0,54         0,94
 Razón de verosimilitud +          8,16        5,51        12,10
```

Razón de verosimilitud -	0,17	0,05	0,61

Aparecen todos los índices mencionados con anterioridad con sus respectivos intervalos de confianza. Es de notar que el usuario (investigador) debe conocer qué diseño empleó para obtener su tabla de datos. Si se trata del primero de los diseños mencionados en el acápite correspondiente, ya se vio que no es posible obtener una estimación adecuada de los VP. Si se trata del segundo diseño, no serán apropiadas las estimaciones de S y de E, y solo si se trata del tercer diseño se podrá hacer un uso apropiado de toda la información que brinda la tabla de resultados de Epidat. Ni la RV ni el índice de Youden podrán estimarse a partir del segundo diseño que solo serán válidos si se obtienen con un diseño como el primero o el tercero.

Si se tuviera el problema de estimar el VP de este test en otro contexto, es decir, en un sitio o ámbito donde la prevalencia de la enfermedad fuera distinta, en tal caso se haría uso de la otra opción que aparece en el submódulo de Pruebas Simples: "valores predictivos". Supóngase, por ejemplo, que interesa conocer el VP del cuestionario en una clínica Psiquiátrica orientada hacia los desórdenes alimentarios, y se tiene el conocimiento de que el 40% de las mujeres que acuden tienen realmente un desorden alimentario. Entonces, se pondría en las casillas correspondientes a la sensibilidad y la especificidad los valores de 85 y 90, respectivamente, y en la casilla de la prevalencia se pondría 40%. Al pedir el cálculo, se obtendría la siguiente tabla:

Pruebas diagnósticas simples

Sensibilidad:	85,00%		
Especificidad:	90,00%		
Prevalencia:	40,00	x	100

	Valor
Indice de Validez (%)	88,00
Valor predictivo + (%)	85,00
Valor predictivo - (%)	90,00
Indice de Youden	0,75
Razón de verosimilitud +	8,50
Razón de verosimilitud -	0,17

donde se observan las estimaciones de los VP para ese contexto. Se obtienen además indicadores que se calculan a partir de la S y de la E y que se tienen también en el cuadro anterior, solo que esta vez se emplea valores aproximados para la S y la E, ya que estaban "actuando" como valores conocidos de la prueba, y los estimadores difieren ligeramente de los obtenidos con la tabla de datos.

PRUEBAS MÚLTIPLES

El uso de pruebas múltiples es muy frecuente en la práctica médica. Ante una, o más de una, sospecha diagnóstica, el médico suele tener varias posibilidades de pruebas que lo ayuden a confirmar o descartar su diagnóstico. Se puede suponer que con más de una prueba se llegará a un diagnóstico más certero. El problema es, entonces, evaluar si tal suposición se cumple, y éste puede ser el objetivo de una investigación. Hay por lo menos dos formas de indicar varias pruebas:

En paralelo. Todas se aplican simultáneamente a la misma muestra de individuos, de forma que se consideran negativos aquellos sujetos que obtienen resultados negativos en todas las pruebas, y positivos todos los demás.

En serie: Se aplica una prueba en primer lugar, y después se indica la otra prueba solo si el individuo resulta positivo de la anterior. Al final, se considera positivo al sujeto que haya tenido resultados positivos en todas las pruebas y negativos a todos los demás.

La sensibilidad y la especificidad global de las pruebas se estiman como hasta ahora, solo que con el resultado global de todas las pruebas.

Ejemplo

Supóngase que se tienen 20 pacientes con cierta dolencia (enfermos verdaderos) y 10 personas en los que se ha comprobado que no tienen la enfermedad. Se desea conocer la eficacia de dos pruebas P1 y P2 aplicadas en paralelo y en serie.

En paralelo se obtienen los siguientes resultados:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba P1	**Positivos**	15	3	18
	Negativos	5	7	12
	Total	20	10	30

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba P2	**Positivos**	12	4	16
	Negativos	8	6	14
	Total	20	10	30

Para poder calcular la sensibilidad y la especificidad de la prueba, hay que conocer quiénes resultaron negativos con las dos pruebas en ambos grupos, pero esto no hay manera de deducirlo de las dos tablas anteriores. En este caso, supóngase que se conoce que, de los negativos en la primera prueba, 2 de los enfermos y 3 de los sanos tuvieron un resultado negativo con la segunda prueba. Esto es, que con los 12 negativos de la primera prueba se podría construir la siguiente tabla:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba P2	**Positivos**	3	4	7
	Negativos	2	3	5
	Total	5	7	12

Se tiene, entonces, la tabla global:

	Criterio de verdad

		Enfermos	No enfermos	Total
Prueba global	**Positivos**	18	7	16
	Negativos	2	3	14
	Total	20	10	30

y de aquí se calculan la S y la E de las dos pruebas conjuntas aplicadas en paralelo.

Trabajando con Epidat 3.1 se marca la opción "en paralelo" y se le indica el número de pruebas que se están evaluando con esos sujetos. Se llenan entonces los valores para las dos tablas que aparecen en el recuadro de la derecha, recordando que la segunda tabla habrá de contener solo los negativos de la primera, igual a como se vio antes. El resultado que se tiene es el siguiente:

```
Pruebas diagnósticas múltiples

          Tipo de pruebas:   En paralelo
        Número de pruebas:            2
       Nivel de confianza:        95,0%

Clasificación final de los sujetos

            Enfermos       Sanos       Total
---------- ---------   ---------   ---------
  Positivo        18           7          25
  Negativo         2           3           5
---------- ---------   ---------   ---------
     Total        20          10          30


                              Valor            IC (95%)
-------------------------    ----------  ------------------------
          Sensibilidad(%)         90,00        74,35       100,00
         Especificidad(%)         30,00        -3,4        63,4
     Indice de Validez(%)         70,00        51,94       88,06
    Valor predictivo +(%)         72,00        52,40       91,60
    Valor predictivo -(%)         60,00        7,06        100,00
           Prevalencia(%)         66,67        48,13       85,20

         Indice de Youden          0,20        -0,11       0,51
 Razón de verosimilitud +          1,29         0,84       1,98
 Razón de verosimilitud -          0,33         0,07       1,68
```

Se observa la tabla global calculada por Epidat, así como los resultados de la S, la E y el resto de los indicadores vistos con anterioridad con sus respectivos intervalos de confianza.

Si ambas pruebas se realizan en serie, entonces se tiene la tabla original correspondiente a P1:

		Criterio de verdad		
		Enfermos	**No enfermos**	**Total**
Prueba P1	**Positivos**	15	3	18
	Negativos	5	7	12
	Total	20	10	30

y el resultado de aplicar la prueba P2 a los 18 clasificados como positivos por P1:

	Criterio de verdad

		Enfermos	No enfermos	Total
Prueba P2	**Positivos**	10	1	11
	Negativos	5	2	7
	Total	15	3	18

Estas son las tablas que se deben introducir en Epidat para obtener el resultado siguiente:

```
Pruebas diagnósticas múltiples

          Tipo de pruebas:     En serie
        Número de pruebas:            2
       Nivel de confianza:         95,0%

Clasificación final de los sujetos

            Enfermos       Sanos       Total
---------- ---------   ---------   ---------
  Positivo        10           1          11
  Negativo        10           9          19
---------- ---------   ---------   ---------
     Total        20          10          30


                              Valor            IC (95%)
-------------------------    ----------  -------------------------
          Sensibilidad(%)         50,00       25,59        74,41
         Especificidad(%)         90,00       66,41       100,00
     Indice de Validez(%)         63,33       44,42        82,24
    Valor predictivo +(%)         90,91       69,38       100,00
    Valor predictivo -(%)         47,37       22,29        72,45
           Prevalencia(%)         66,67       48,13        85,20

         Indice de Youden          0,40        0,11         0,69
 Razón de verosimilitud +          5,00        0,74        33,78
 Razón de verosimilitud -          0,56        0,34         0,90
```

donde se observa la tabla final (global) construida por Epidat 3.1 y todos los indicadores.

Este submódulo también permite calcular los valores predictivos de pruebas en serie o en paralelo, siempre que se conozca la prevalencia de la enfermedad y la S y E de cada una de las pruebas que se usen en serie o en paralelo.

Supóngase que las pruebas del ejemplo anterior quieren utilizarse en un contexto donde la prevalencia de la enfermedad en cuestión es del 10%. ¿Qué valor predictivo tendría entonces la combinación de las dos pruebas? Se conoce que la S de P1 es 75%, la E de P1 es 70% y la S y la E de P2 son ambas de 60%.

En serie:

```
Pruebas diagnósticas múltiples

          Tipo de pruebas:     En serie
        Número de pruebas:            2
              Prevalencia:        10,00 x      100
```

	Valor
Sensibilidad (%)	45,00
Especificidad (%)	88,00
Indice de Validez (%)	83,70
Valor predictivo + (%)	29,41
Valor predictivo - (%)	93,51
Indice de Youden	0,33
Razón de verosimilitud +	3,75
Razón de verosimilitud -	0,63

En paralelo:

Pruebas diagnósticas múltiples

Tipo de pruebas: En paralelo
Número de pruebas: 2
Prevalencia: 10,00 x 100

	Valor
Sensibilidad (%)	90,00
Especificidad (%)	42,00
Indice de Validez (%)	46,80
Valor predictivo + (%)	14,71
Valor predictivo - (%)	97,42
Indice de Youden	0,32
Razón de verosimilitud +	1,55
Razón de verosimilitud -	0,24

PRUEBAS DE REFERENCIA IMPERFECTAS

Hasta ahora se ha venido trabajando con la idea de que evaluar la eficacia de una PD transita por el conocimiento de la verdad, o lo que es lo mismo, por la existencia de alguna manera, independiente de la prueba, de arribar al diagnóstico verdadero de los pacientes que son incluidos en el estudio.

Sin embargo, frecuentemente no existe una manera viable de arribar a la verdad, bien porque no puede realizarse una exploración invasiva por razones éticamente indiscutibles, o bien porque la tal prueba de la verdad no puede realizarse en un límite de tiempo razonable. En tal caso, afortunadamente, se pueden dar dos situaciones prácticas a las que se les ha encontrado una solución matemática que conduce a estimadores adecuados de los indicadores básicos de la prueba en estudio.

En la primera situación se cuenta con una prueba de referencia imperfecta cuya sensibilidad y especificidad se conocen. En el caso de que se tenga un diseño con una muestra de N pacientes (o sujetos) a los que se les han aplicado ambas pruebas, la de referencia y la nueva se puede demostrar que[8]:

$$S = \frac{(a+b)\beta - b}{(a+c) - (1-\beta)N} \qquad E = \frac{(c+d)\alpha - c}{N\alpha - (a+c)}$$

donde α es la sensibilidad de la prueba de referencia, β es su especificidad, y *a, b, c* y *d* los símbolos para las celdas empleados desde el inicio (ver Tabla 1). S y E brindan los estimadores respectivos de sensibilidad y especificidad de la prueba nueva.

Igualmente se puede estimar la prevalencia en ese contexto como:

$$p = \frac{\frac{(a+c)}{N} + \beta - 1}{\alpha + \beta - 1}$$

y, dado que se tiene una muestra N de sujetos que van a ser evaluados con ambas pruebas, también se pueden calcular los valores predictivos de la PD en ese contexto.

Además también se calculan los intervalos de confianza para todos estos valores ajustados, por el método Bootstrap mediante la técnica del percentil simple (Efron, 1993).

Ejemplo

Se tiene una prueba nueva, para la cual se quieren calcular los indicadores de eficacia. No se tiene un criterio de verdad disponible ni sujetos en los que se conozca el verdadero diagnóstico por otras vías, pero se tiene otra prueba con S=0,9 y E=0,6 que puede servir como prueba de referencia.

Los resultados de la tabla de 2x2 que surge de evaluar a los sujetos con ambas pruebas se muestran a continuación:

		Prueba de referencia (S=90% y E=60%)		
		Enfermos	**No enfermos**	**Total**
Prueba Nueva	**Positivos**	84	26	110
	Negativos	46	44	90
	Total	130	70	200

Cuando se introducen los datos en Epidat 3.1 se obtiene lo siguiente:

```
Prueba de referencia imperfecta

Prueba de referencia
             Sensibilidad:        90,00%
            Especificidad:        60,00%

                      Prueba de referencia
  Prueba diagnóstica    Positivo    Negativo       Total
--------------------   ---------   ---------   ---------
            Positivo          84          26         110
            Negativo          46          44          90
--------------------   ---------   ---------   ---------
               Total         130          70         200



     RESULTADOS AJUSTADOS         Valor      IC (95%) Bootstrap
```

Sensibilidad (%)	80,00	65,09	98,57
Especificidad (%)	70,00	55,53	85,26
Índice de validez (%)	75,00	62,50	87,50
Valor predictivo + (%)	72,73	56,29	88,33
Valor predictivo - (%)	77,78	56,92	98,60
Prevalencia (%)	50,00	37,00	63,00
Índice de Youden	0,50	0,25	0,77
Razón de verosimilitud +	2,67	1,59	5,76
Razón de verosimilitud -	0,29	0,02	0,58

RESULTADOS SIN AJUSTAR	Valor	IC (95%)	
Sensibilidad (%)	64,62	56,01	73,22
Especificidad (%)	62,86	50,82	74,89
Índice de validez (%)	64,00	57,10	70,90
Valor predictivo + (%)	76,36	67,97	84,76
Valor predictivo - (%)	48,89	38,01	59,77
Prevalencia (%)	65,00	58,14	71,86
Índice de Youden	0,27	0,13	0,41
Razón de verosimilitud +	1,74	1,25	2,42
Razón de verosimilitud -	0,56	0,42	0,76

Una segunda posibilidad consiste en aplicar más de una vez la prueba en estudio a los mismos sujetos. En este caso, debe suponerse que la prueba arroja resultados consistentes, es decir, siempre dará el mismo resultado si se aplica al mismo sujeto en iguales condiciones, una suposición acorde con la lógica elemental.

En este caso, se sugiere[8] un proceso iterativo que implica aplicar la prueba, en k ocasiones independientes, a una muestra aleatoria de *n* sujetos de determinada población. El proceso conduce a estimadores máximo verosímiles de S, E y la prevalencia (o probabilidad *a priori*).

Epidat 3.1 incorpora también este procedimiento. Los datos que hay que proporcionarle al sistema para que desarrolle el procedimiento iterativo son los siguientes:

- El número n_i de sujetos con i pruebas positivas (i= 1, 2, …, k), que son datos obtenidos del experimento o estudio diseñado para esta estimación.
- El número "estimado" inicial (para el comienzo del proceso iterativo) de sujetos realmente enfermos dentro de los n_i con i resultados positivos.

Ejemplo

Véase un ejemplo con Epidat. Supóngase que se ha decidido (y es factible) realizar 3 veces una prueba a 20 sujetos, cuya S y E desean conocerse en cierto contexto. Después de realizar las tres pruebas los resultados son los siguientes:

- Sujetos con 0 resultado positivo $n_0 = 6$
- Sujetos con 1 resultado positivo $n_1 = 2$
- Sujetos con 2 resultados positivos $n_2 = 6$
- Sujetos con 3 resultados positivos $n_3 = 6$

Para poder obtener un estimador máximo verosímil de la S y la E de esta prueba (sin criterio de verdad) se debe "inventar" un número inicial de sujetos realmente enfermos dentro de cada una de las n_i, por ejemplo:

- Sujetos "inventados" realmente enfermos dentro de $n_0 = 1$
- Sujetos "inventados" realmente enfermos dentro de $n_1 = 1$
- Sujetos "inventados" realmente enfermos dentro de $n_2 = 4$
- Sujetos "inventados" realmente enfermos dentro de $n_3 = 6$

Entonces a Epidat, en la opción *Sensibilidad y especificidad desconocidas* del submódulo de "Prueba de referencia imperfecta" se le darán los datos anteriores en la columna "Enfermos", y en la columna de "Total" se le introducen los verdaderos datos que son las n_i.

El número de enfermos debe ser menor que el número total de sujetos; si en algún caso se introduce un dato mayor en la columna "Total" no se activa la calculadora.

Véase la salida de Epidat:

```
Prueba de referencia imperfecta

        Número de pruebas:            3

Prueba diagnóstica que se evalúa

                                   Valor
-------------------------    ----------
         Sensibilidad (%)         78,82
        Especificidad (%)         94,29
          Prevalencia (%)         65,14

    Indice de Validez (%)         84,21
   Valor predictivo + (%)         96,27
   Valor predictivo - (%)         70,44

         Indice de Youden          0,73
 Razón de verosimilitud +         13,80
 Razón de verosimilitud -          0,22
```

LAS PRUEBAS CON MÁS DE DOS RESULTADOS

Curva ROC

Hasta el momento se ha hablado de pruebas que son aplicadas a dos grupos de la población, el grupo **con** y el grupo **sin** la enfermedad. Los resultados de tales pruebas son citados como positivos o negativos según señale o no hacia la presencia de la enfermedad en cuestión. Pero la realidad suele ser más compleja que los modelos que el hombre busca para representarla. En algunas instancias, más de dos categorías pueden ser necesarias para enmarcar la condición de cada paciente, el resultado de una prueba, o de ambos.

Uno de estos casos es cuando los resultados de una prueba son de naturaleza cuantitativa u ordinal, o sea, el resultado de realizar el test diagnóstico es un número, un rango, o un nivel (v.g. 3,4 mmol/L, "ligero", 36 puntos, etc.), y es el médico el que decide cuál es el punto del espectro cuantitativo (o semicuantitativo) que permite separar a los enfermos de los no enfermos. Hay que recordar que el médico tendrá siempre que decidir dicotómicamente (tratar o no), pero es obvio que, en estos casos, la decisión es equivalente a señalar un punto, en el rango de resultados posibles, que divide a los paciente en probablemente enfermos y probablemente no enfermos. De modo que, para conocer la eficacia de una prueba de este tipo, habrá que decidir el punto de corte (PC) que permita declarar a las personas con resultado positivo o negativo, y estimar entonces los indicadores de eficacia según se ha visto en las secciones anteriores. Está claro que las estimaciones de sensibilidad y especificidad de una prueba como ésta dependerán del punto de corte seleccionado, y que el médico deberá escoger el punto de corte óptimo según sus necesidades. La selección de un punto de corte óptimo es, con este tipo de pruebas, la tarea más importante. Sin embargo, la noción de PC óptimo no es única ya que, por un lado, son casi inexistentes los tests con S y E ambas muy altas (cerca de 1) y, por otro lado, la práctica clínica es versátil en sus necesidades de S y E altas.

El siguiente ejemplo, tomado de Feinstein[18], ilustra la situación cuando cambia el PC para una prueba dada. Se trata de dos grupos de pacientes, uno de los cuales tiene una enfermedad coronaria (EC) demostrada y otro que no la tiene. A todos los pacientes se les realizó la prueba ergométrica y se les midió el desnivel del segmento ST al final de la prueba.

Desnivel de ST	Con EC	Prop. acumulada	Sin EC	Prop. acumulada
≥ 3mm	31	0,21	0	0,00
2,5-<3,0	15	0,31	0	0,00
2,0-<2,5	27	0,49	7	0,05
1,5-<2,0	30	0,69	8	0,10
1,0-<1,5	32	0,90	39	0,36
0,5-<1,0	12	0,98	43	0,65
<0,5	3	1,00	53	1,00
Total	**150**		**150**	

Se observa en la tabla, por ejemplo, que la proporción de individuos enfermos con un desnivel mayor de 2 mm es de 0,49, mientras que la de individuos no enfermos es de 0,05. En términos de los indicadores vistos en temas anteriores, para este punto de corte la sensibilidad de la prueba ergométrica, o proporción de enfermos positivos (desnivel > 2mm), se estima en 0,49 ó 49% y la especificidad, o proporción de no enfermos negativos, en 0,95 (1-0,05) ó 95%. Se hace evidente también que un cambio en el punto de corte genera un cambio en la S y la E de la prueba y que, en

este caso, donde un desnivel mayor del segmento ST indica "mayor anormalidad", a medida que el punto de corte aumenta en valor, aumenta la E y disminuye la S.

Para una prueba diagnóstica cuyo resultado es cuantitativo, es entonces imposible estimar indicadores de eficacia a menos que se señale un punto de corte determinado. Surge aquí la necesidad de encontrar un indicador general de eficacia para este tipo de pruebas. La llamada curva ROC brinda el indicador necesario.

Esta curva fue por primera vez propuesta en el decenio de los 50 para describir la relación entre señal y ruido, y se desarrolló en la comparación de la eficacia de radares. Se necesitaba evaluar la capacidad de un radar para distinguir entre verdaderas señales y ruido de otros tipos. El radar podría equivocarse de dos formas: fallando en la detección de la señal (falso negativo) o detectando una falsa (falso positivo). A los radares se les cambia el umbral de detección de señales y este cambio origina distintas tasas de errores relacionados entre sí: a medida que el umbral disminuye, la tasa de falsos negativos desciende (aumenta la sensibilidad) y aumenta la tasa de falsos positivos (disminuyendo la especificidad). Las siglas ROC vienen de su nombre en inglés: *Receiver Operating Characteristic Curve,* que se traduce como *Curva de Características Operacionales del Receptor*.

Estas curvas se trasladaron con facilidad a la evaluación de PD, ya que la situación es muy similar. Se trata de detectar una enfermedad dada y la PD puede equivocarse en el sentido de los falsos negativos o los falsos positivos; al cambiar el punto de corte (homólogo del umbral para el radar) cambian las tasas de error (o la S y E, si se prefiere).

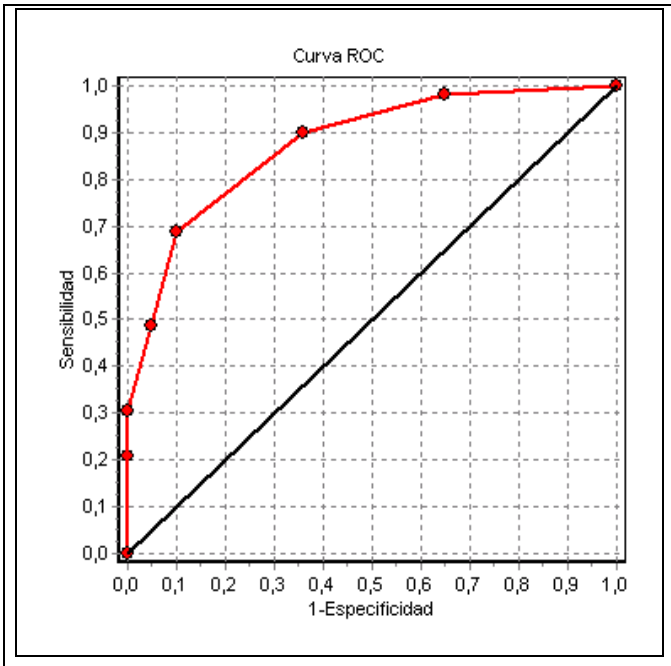
La curva ROC empírica típica se construye al representar en dos ejes de coordenadas los puntos (x, y) dados por (1-E, S) para cada punto de corte. Véase cómo se obtiene con Epidat 3.1 la curva ROC correspondiente al ejemplo visto anteriormente.

En el submódulo de Curvas ROC simples Epidat ofrece 3 opciones para introducir los datos: 1) Tablas 2x2, 2) Tablas Kx2 y 3) Datos continuos. Tal como aparecen los datos anteriores la mejor opción es la 2. En este caso, Epidat permite introducir las categorías de datos agrupados y el número de enfermos y no enfermos en cada categoría. Epidat 3.1 exige, para representar adecuadamente la curva ROC por encima de la diagonal, que las categorías estén ordenadas de menor a mayor "anormalidad". Por tanto, tal y como están en la tabla los datos del ejemplo, deben introducirse empezando por la última fila (categoría: <0,5 mm) e ir "subiendo" hacia la primera (categoría: > 3 mm).

Véanse los resultados:

```
Curvas ROC simples

     Número de categorías:            7
       Nivel de confianza:         95,0%
```



Area ROC	EE	IC(95%)		
0,8753	0,0189	0,8382	0,9125	Delong
	0,0205	0,8351	0,9156	Hanley & McNeil

Epidat también brinda la posibilidad de introducir los datos mediante tablas 2x2 para cada punto de corte (opción 1). En este caso serían 6 tablas de 2x2 construidas de la siguiente forma:

Tabla 1

	Criterio de verdad Punto de corte: 3 mm	
	Con EC	Sin EC
Positivos	31	0
Negativos	119	150
Total	150	150

Tabla 2

	Criterio de verdad Punto de corte: 2,5 mm	
	Con EC	Sin EC
Positivos	46	0
Negativos	104	150
Total	150	150

Tabla 3

	Criterio de verdad Punto de corte: 2 mm	
	Con EC	Sin EC
Positivos	73	7
Negativos	77	143
Total	150	150

Tabla 4

	Criterio de verdad Punto de corte: 1,5 mm	
	Con EC	Sin EC
Positivos	103	15
Negativos	47	135
Total	150	150

Tabla 5

	Criterio de verdad Punto de corte: 1 mm	
	Con EC	Sin EC
Positivos	135	54
Negativos	15	96
Total	150	150

Tabla 6

	Criterio de verdad Punto de corte: 0,5 mm	
	Con EC	Sin EC
Positivos	147	97
Negativos	3	53
Total	150	150

Los resultados obtenidos de introducir las 6 tablas anteriores en Epidat 3.1 son los mismos que en el caso anterior.

La tercera forma de introducir los resultados en Epidat es mediante el valor individual obtenido en la prueba para cada individuo. En esta opción se da la posibilidad de introducir los datos crudos o de importar la base de datos con los datos originales. En cualquier caso habrá que comunicar quiénes son los verdaderos enfermos y quiénes no.

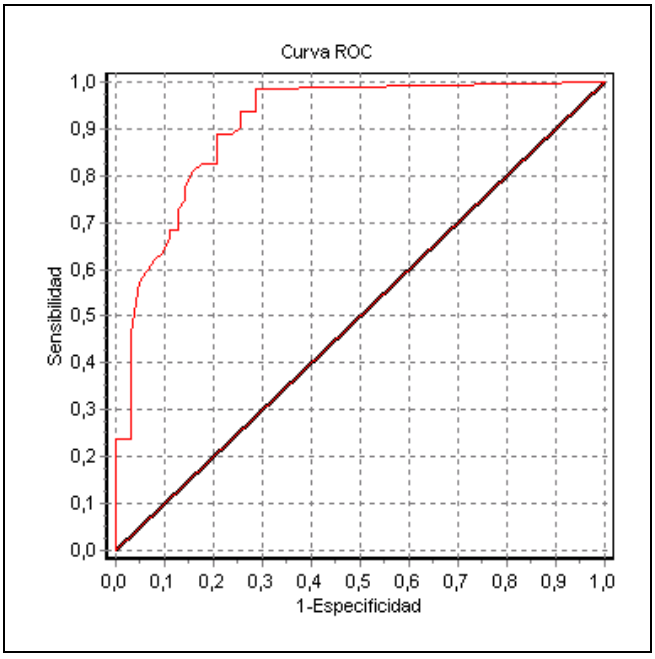
Ejemplo

Véase un ejemplo con una base de datos sobre el valor de cierto índice que se obtiene en la ultrosonografía doppler de mama. Se tienen 63 mujeres con cáncer de mama y 63 mujeres que no padecen la enfermedad, y de todas se tiene el valor del índice. Los datos se encuentran en una base de datos Excel, llamada RUDDY INDICE.xls, con 2 campos: INDICE y GRUPO. El grupo indica la enfermedad y se le puso el código 0 a las sanas y el 1 a las enfermas. Al cargar los datos en Epidat se obtienen los resultados siguientes:

```
Curvas ROC simples

Archivo de trabajo: C: \Archivos de programa \Epidat 3.1 \Ejemplos
\Pruebas diagnósticas \RUDDY INDICE.xls
Campo que identifica:
   Enfermedad:  GRUPO
   Resultado de la prueba:  INDICE

        Número de enfermos:          63
           Número de sanos:          63
        Nivel de confianza:       95,0%
```



Area ROC	EE	IC(95%)		
0,9101	0,0258	0,8595	0,9606	Delong
	0,0271	0,8569	0,9632	Hanley & McNeil

Como se ha visto, en todos los casos Epidat brinda el gráfico con la curva ROC y la estimación del área bajo la curva con error estándar (EE) y su intervalo de confianza calculados mediante dos métodos[22,23].

El área bajo la curva tendrá un valor menor que 1, y constituye el indicador de eficacia más general para pruebas de este tipo. Se puede demostrar que una PD que tenga un punto de corte tal que S y E sean altas (dígase mayores de 0,8) tendrá una curva con concavidad más afilada, y la curva ROC de una prueba perfecta (S y E iguales a 1) será las líneas formadas por los propios ejes izquierdo y superior del recuadro que sirve de marco a la curva.

Comparación de curvas ROC

Es natural que la comparación de la eficacia de dos o más pruebas diagnósticas para detectar una enfermedad o proceso patológico dado, pueda hacerse sobre la base de comparar los valores de S y de E de tales pruebas. Pero, cuando se trata de PD con resultado cuantitativo, la comparación de las curvas ROC correspondientes resulta el modo más natural de determinar cuál de las pruebas es más eficaz, puesto que ya se vio que la S y la E de pruebas de este tipo depende del punto de corte que se elija. Teniendo en cuenta lo que se ha visto hasta ahora se comprende que la curva ROC que tenga el área mayor será la que corresponde a la prueba más eficaz.

Con Epidat 3.1 se pueden hacer comparaciones de curvas ROC[22]. Hay dos maneras de introducir los datos para lograr la comparación necesaria: 1) cuando se tienen los datos en tablas Kx2 y 2) cuando se tienen los datos continuos para cada individuo.

Cuando se tienen los datos en tablas, se supone que haya una tabla para cada prueba, que todas las tablas tengan igual número de categorías y que todas las pruebas se les hayan aplicado a los mismos individuos (sanos y enfermos). Se especifica el número de categorías en el lugar indicado en el sistema y debe decidirse si la entrada se hará de forma manual (directamente en Epidat desde el teclado) o automática (a partir de una base de datos en un fichero aparte). En el primer caso, aparece, en el sitio donde deben introducirse los datos, una tabla con espacios de la manera que se ilustra a continuación, suponiendo que se comparan dos curvas con 4 categorías. Si se tuviera que comparar 3 curvas, la tabla de entrada de datos tendría 4 filas más.

Categoría	Curva	Enfermos	Sanos
1	1		
2	1		
3	1		
4	1		
1	2		
2	2		
3	2		
4	2		

En esta tabla deberemos incluir los datos de la primera prueba (curva=1) y de la segunda prueba (curva=2).

Ejemplo

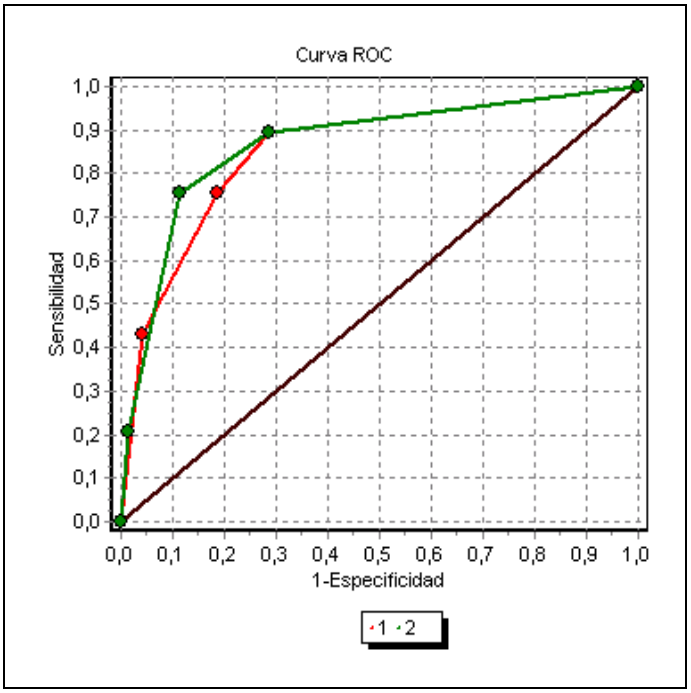
Veamos un ejemplo de comparación de las curvas de dos pruebas con 4 categorías cada una. Se trata de comparar la eficacia del "número de vasos" que se obtiene con el US-doppler a color de un nódulo de mama para el diagnóstico de malignidad y el valor del índice que se obtiene con el mismo procedimiento. La prueba 1 es el número de vasos con 4 categorías (1 = 0 vasos; 2 = 1-3 vasos; 3 = 4-6 vasos y 4 = más de 6 vasos) y la prueba 2 es el índice, también con 4 categorías (1 = 0; 2 = 0,01-0,38; 3 = 0,39-0,63 y 4 = >0,63). Los datos se muestran en la siguiente tabla:

	PRUEBA 1		PRUEBA 2	
Categoría	**Enfermos**	**Sanos**	**Enfermos**	**Sanos**
1	8	50	8	50
2	11	7	11	12
3	25	10	42	7
4	33	3	16	1

Al introducir estos datos en Epidat se obtiene el siguiente resultado:

```
Comparación de curvas ROC

          Número de curvas:            2
     Número de categorías:             4
       Nivel de confianza:         95,0%
```



Curva	Area ROC	EE	IC(95%)	
1	0,8532	0,0306	0,7933	0,9132
2	0,8681	0,0292	0,8109	0,9252

Prueba de homogeneidad de áreas

Ji-cuadrado	gl	Valor P
0,1234	1	0,7254

Se observa que ambas pruebas tienen una eficacia semejante en el gráfico y que sus áreas son prácticamente iguales.

Si los datos se introducen de forma automática, a partir de una base de datos en Excel, habrá que indicarle al sistema dónde está disponible la base de datos (se tiene la posibilidad de examinar el entorno de la PC e indicarle el fichero en el directorio que se encuentre); además, los datos deben estar organizados de forma que una curva esté a continuación de la otra. Con los datos del ejemplo, se tendría la siguiente disposición en un fichero Excel (se incluye en Epidat 3.1 con el nombre MAMA.xls):

CATEGORÍA	CURVA	ENFERMOS	SANOS
1	1	8	50
2	1	11	7
3	1	25	10
4	1	33	3
1	2	8	50
2	2	11	12
3	2	42	7
4	2	16	1

Una tabla estaría a continuación de la otra. A Epidat habría que decirle el campo donde está la curva ("CURVA", en este caso), donde están los enfermos ("ENFERMOS", en este caso) y donde están los sanos ("SANOS", en este caso). Los resultados serían los mismos que en el caso anterior, cuando la entrada se hizo de forma manual.

Si en lugar de tablas Kx2, se tienen datos continuos, Epidat permite seleccionar entre lo que llama curvas independientes (pruebas aplicadas a distintos individuos) o curvas correlacionadas (pruebas aplicadas a los mismos individuos). También hay dos opciones para la entrada de datos, una manual y otra automática.

Si se trata de curvas independientes y la entrada de datos se realiza de forma manual, se le debe informar al sistema el número de curvas y el número de enfermos y sanos en cada curva (es de suponer que sean distintos pues son pruebas independientes). A continuación, debe introducirse el valor de la prueba para cada uno de los sujetos de las diferentes curvas. Si los datos se tuvieran en una base de datos Excel para entrada automática, esta base debe tener 3 campos por lo menos, uno que identifique la curva, uno que identifique la enfermedad (1 = Sí, 0 = No), y otro que identifique el valor de las pruebas.

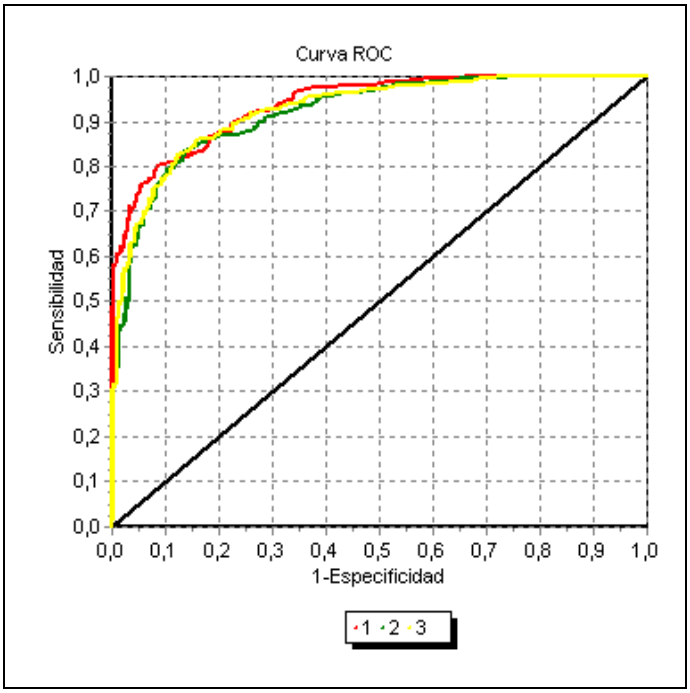
Ejemplo

Véase un ejemplo de datos continuos con tres pruebas independientes. Los datos se encuentran en el archivo PRUEBAS-INDEP.xls, incluido en Epidat 3.1 y, al cargarlos en el programa, se obtienen los siguientes resultados:

```
Comparación de curvas ROC

Archivo de trabajo: C: \Archivos de programa \Epidat 3.1 \Ejemplos
\Pruebas diagnósticas \PRUEBAS-INDEP.xls
Campo que identifica:
   Identificación de curva :  CURVA
   Número de enfermos :  ENFERMEDAD
   Número de sanos :  PRUEBA

          Tipo de curvas:   Curvas independientes
        Número de curvas:            3
     Nivel de confianza:          95,0%
```



Curva	Area ROC	EE (DeLong)	IC(95%)	
1	0,9380	0,0081	0,9221	0,9539
2	0,9192	0,0103	0,8990	0,9395
3	0,9251	0,0099	0,9056	0,9446

Prueba de homogeneidad de áreas

Ji-cuadrado	gl	Valor P
2,2701	2	0,1319

Se trata de 3 pruebas con eficacia similar, lo que se refleja en el gráfico y en las áreas bajo las curvas.

Cuando las curvas son correlacionadas, lo que debe indicársele al sistema es el campo que tiene la "enfermedad" (1 = Sí, 0 = No), y los campos que tienen los resultados de las diferentes pruebas.

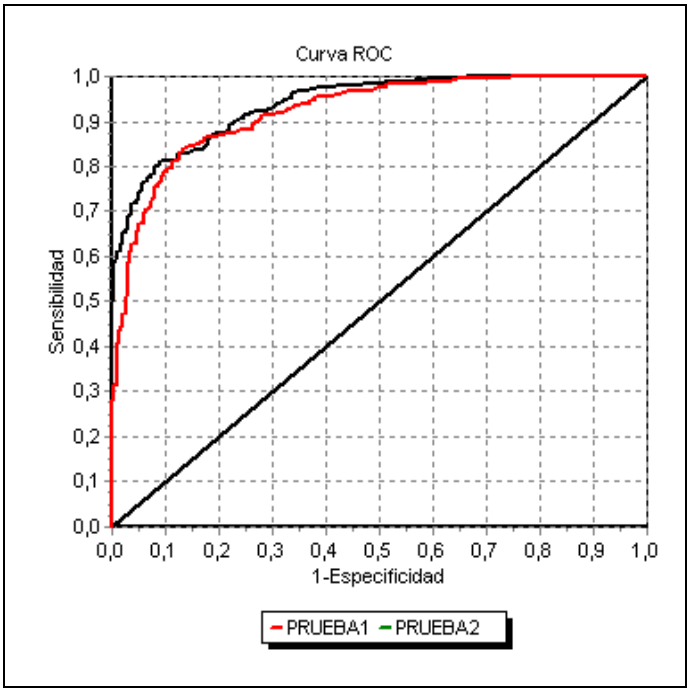
Ejemplo

Véanse los resultados de Epidat de un ejemplo de comparación de dos pruebas correlacionadas, o sea, aplicadas a los mismos sujetos sanos y enfermos. Los datos se encuentran en el archivo PRUEBAS CORREL.xls.

```
Comparación de curvas ROC

Archivo de trabajo: C: \Archivos de programa \Epidat 3.1 \Ejemplos
\Pruebas diagnósticas \PRUEBAS CORREL.xls
Campo que identifica:
   Número de enfermos :  ENFERMEDAD

          Tipo de curvas:   Curvas correlacionadas
        Número de curvas:            2
     Nivel de confianza:          95,0%
```



Curva	Area ROC	EE (DeLong)	IC(95%)	
PRUEBA2	0,9390	0,0079	0,9235	0,9544
PRUEBA1	0,9213	0,0099	0,9018	0,9408

Prueba de homogeneidad de áreas

Ji-cuadrado	gl	Valor P
2,1589	1	0,1417

Se nota también que ambas curvas son similares, lo que significa que ambas pruebas tienen una eficacia similar para el diagnóstico de la enfermedad en cuestión.

LA CURVA DE LORENZ

La curva ROC y sus índices son útiles cuando el riesgo de enfermedad aumenta o disminuye de forma monótona con los valores de la prueba. Cuando el riesgo no es monótono, la curva ROC que resulta puede no ser convexa y sus índices no fiables. Para estas situaciones, Lee[24] propuso una alternativa basada en curvas similares a las de Lorenz y los índices de Pietra y de Gini.

La curva de Lorenz, dicho someramente, es un instrumento gráfico del área de la economía, desarrollado por Mark O. Lorenz para representar las desigualdades en los ingresos de los hogares (personas, grupos, etc.) en cierta región.

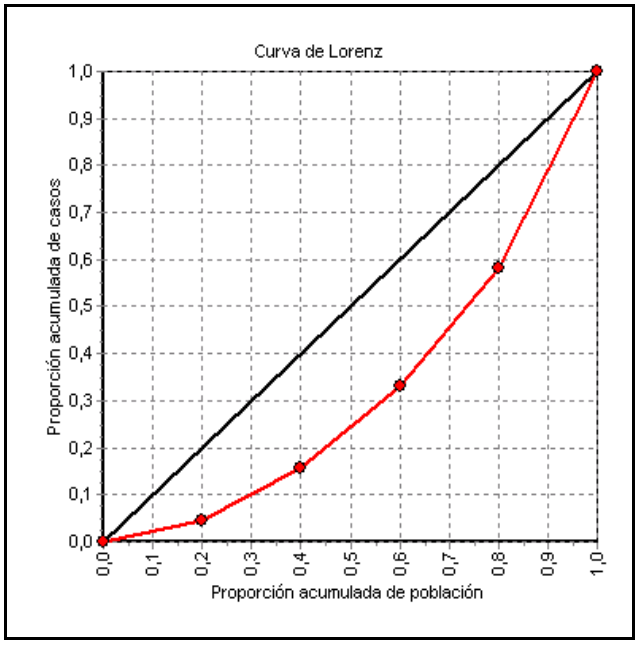
Se trata de un gráfico de coordenadas; en el eje de las X se representa el porcentaje acumulado de hogares, mientras que en el eje de las Y se representa el porcentaje acumulado de ingresos. Véase un ejemplo. Supóngase que se tienen los datos de ingresos anuales de los hogares de determinada población. Esos datos se ordenan de menor a mayor y, por ejemplo, se obtienen los 5 quintiles de esa serie ordenada. El primer quintil es una cantidad dada que se puede expresar en porcentaje del total de ingresos y así todos los quintiles se expresan de esa forma. Como se trata de quintiles se sabe que el 20% de la población estará por debajo del primer quintil, el 40% debajo del segundo quintil, y así sucesivamente. Con estos datos se construye la curva de Lorenz.

La siguiente tabla representa un ejemplo:

% acumulado de hogares	% de ingresos	% acumulado de ingresos
0	0	0
20	4,7	4,7
40	11	15,7
60	17,4	33,1
80	25	58,1
100	41,9	100

Se observa que el 20% de la población de hogares tiene solo el 4,7% de los ingresos y que el 60% de los hogares acumulan el 33,1% de los ingresos. Se obtiene una manera de medir la desigualdad imperante en esa población, puesto que si los ingresos estuvieran igualmente distribuidos, el 20% de la población tendría el 20% de los ingresos y así sucesivamente.

La representación de estos datos en un gráfico como el que se describió da la llamada curva de Lorenz:



Se observa que la concavidad de la curva da una idea de la magnitud de las desigualdades, y que la igualdad perfecta se reflejaría en una curva como la diagonal del cuadrado que forman los ejes. De este modo, el área comprendida entre la curva y la diagonal puede dar una idea de la magnitud de las desigualdades. El índice de Gini no es más que esa área expresada como proporción del área total debajo de la curva de igualdad perfecta (la diagonal del cuadrado). Se puede obtener, además, el índice de Pietra, que es el área del mayor triángulo inscrito entre la curva de Lorenz y la diagonal.

La teoría anterior se puede extrapolar al área de la evaluación de Pruebas Diagnósticas. Cuando se tiene una prueba con resultado cuantitativo se calculan la sensibilidad y la especificidad de la prueba para cada punto de corte seleccionado. Ambas son una especie de proporciones acumuladas, pues cuando los datos se tienen en una tabla Kx2, la sensibilidad se va calculando con la proporción acumulada de positivos a la prueba entre los enfermos a medida que se va cambiando el punto de corte.

Retómese el ejemplo del capítulo para Curva ROC; la prueba era el desnivel ST de la prueba ergométrica y se tenía un grupo de 150 pacientes con enfermedad coronaria comprobada y un grupo de 150 personas comprobadamente sin enfermedad coronaria.

Si se toma como punto de corte el valor de 2 mm en el desnivel del segmento ST (la positividad aumenta con el valor del ST) se tiene que la sensibilidad aquí es la proporción acumulada de pacientes con un nivel de ST mayor que 2 mm en los enfermos (73/150=0,49) y la especificidad la proporción acumulada de personas con el nivel de ST<2 entre los no enfermos (143/150=0,95). Se pueden también calcular las razones de verosimilitud "acumuladas" para cada punto de corte, que son la base para la curva de Lorenz en este ámbito. Para cada razón de verosimilitud en cada punto de corte se tiene la proporción acumulada de enfermos y de no enfermos.

Desnivel de ST	Con EC	Prop. acumulada	Sin EC	Prop. acumulada
> 3 mm	31	0,21	0	0,00
2,5-3,0	15	0,31	0	0,00
2,0-<2,5	27	0,49	7	0,05
1,5-<2,0	30	0,69	8	0,10
1,0-<1,5	32	0,90	39	0,36
0,5-<1,0	12	0,98	43	0,65
<0,5	3	1,00	53	1,00
Total	**150**		**150**	

En resumen, para obtener la curva de Lorenz de una prueba con resultados cuantitativos se siguen los siguientes pasos:

Se ordenan las categorías en función de la razón de verosimilitud, en orden creciente, y se calculan las proporciones acumuladas de enfermos y no enfermos: X_i (proporción acumulada de no enfermos hasta la categoría i) e Y_i (proporción acumulada de enfermos hasta la categoría i). El índice i denota la i-ésima categoría de la prueba diagnóstica (i=1, 2, ..., k), una vez que se reordenaron las categorías de acuerdo a la razón de verosimilitud.

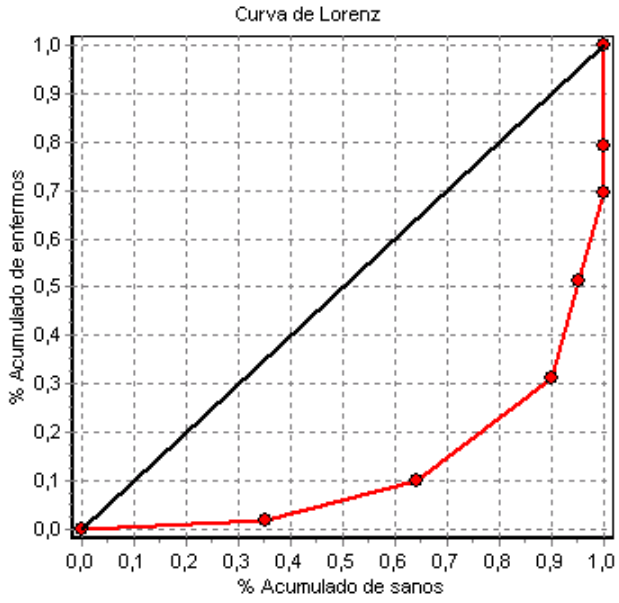
La curva de Lorenz se obtiene representando los puntos (X_i, Y_i) en un cuadrado de lado 1 y uniendo dichos puntos por una línea. Igual que en la curva ROC, se representa la diagonal y los puntos (0,0) y (1,1).

La interpretación de la Curva de Lorenz en el área de las pruebas diagnósticas es menos sencilla que en el área económica, pero también tiene que ver con desigualdades. La curva da una idea de la desigualdad entre S y 1-E (la sensibilidad y el complemento de la especificidad), puesto que si en cada punto de corte la proporción acumulada de enfermos fuera igual a la de no enfermos la sensibilidad y 1-especificidad serían iguales en todos los puntos. En ese caso, la curva sería la diagonal del cuadrado y la RV sería igual a 1 en todos los puntos, lo que denotaría que la PD no tiene valor alguno como medio para diagnóstico. Lo contrario sería que la curva tuviera una concavidad máxima, entonces la PD sería perfecta. De modo que, en el área de la evaluación de PD, a mayor concavidad de la curva de Lorenz mejor es la PD que se está evaluando. Igualmente, un índice de Gini y el de Pietra cercanos a 1 indican bondad de la PD, mientras que si se acercan a 0 denotan la inadecuidad de la PD.

En Epidat 3.1 la construcción de la curva de Lorenz se hace muy fácilmente, introduciendo la tabla del número de enfermos y no enfermos en cada categoría. Para el ejemplo anterior los resultados con Epidat serían los siguientes.

```
Curva de Lorenz

    Número de categorías:            7
```



Curva de Lorenz	
Indice de Gini	0,751
Indice de Pietra	0,587

BIBLIOGRAFÍA

1. Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excesive testing. *N Engl J Med* 1989; 320: 1489-91.

2. Gaarder KR. Diagnosis. *South Med J* 1989; 82: 1153-4.

3. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980; 36: 167-71.

4. Fescina RH, Simini F, Belitzky R. Evaluación de los procedimientos diagnósticos. Aspectos metodológicos. *Salud Perinatal PP* 1985; 2: 39-43.

5. Kassirer JP. Diagnostic Reasoning. *Ann Intern Med* 1989; 110: 893-5.

6. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. *Ann Intern Med* 1981; 94: 557-63.

7. Kassirer JP, Kopelman RI. The luxuriant language of diagnosis. *Hosp Pract* 1989; 24: 36-8.

8. Silva LC. *Métodos estadísticos para la investigación epidemiológica*. Seminario internacional de estadísticas en Euskadi. Instituto Vasco de Estadística; 1987.

9. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; 39: 207-15.

10. Feinstein AR. Clinical biostatistics. XXXI. On the sensitivity, specificity and discrimination of diagnostic tests. *Clin Pharmacol Ther* 1975; 17: 104-16.

11. Taube A. Sensitivity, specificity and predictive values: a graphical approach. *Stat Med* 1986; 5: 585-91.

12. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992; 45: 1143-54.

13. Hlatky MA, Mark DB, Harrell FE, Lee KL, Califf RRM, Pryor DB. Rethinking Sensitivity and Specificity. *Am J Cardiol* 1987; 59: 1195-8.

14. Robertson A, Zweig MH, Van Steirteghem AC. Evaluating the clinical efficacy of Laboratory Tests. *Am J Clin Pathol* 1983; 79(1): 78-86.

15. Sox HC. Probability theory in the use of Diagnostic Tests. *Ann Intern Med* 1986; 104: 60-6.

16. Riegelman RK, Hirsch RP. Discriminación diagnóstica de las pruebas. *Bol Oficina Sanit Panam* 1991; 111: 536-47.

17. Sackett DL, Brian Haynes R, Tugwell P. *Clinical Epidemology: A basic Science for Clinical Medicine*. Madrid: Díaz de Santos; 1989.

18. Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. W.B. Saunders Company; 1985.

19. Gmurman VE. *Teoría de las probabilidades y estadística matemática*. Moscú: Moscú Editorial; 1974.

20. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd Edition. New York: John Wiley and Sons; 1981.

21. Luck AJ, Morgan JF, Reid F, O'Brien A, Brunton J, Price C, et al. The SCOFF questionnaire and clinical interview for eating disorders in general practice: comparative study. *BMJ* 2002; 325: 755-6.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics* 1988; 44: 837-45.

23. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982; 143: 29-36.

24. Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982; 143: 29-36.

25. Lee W. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz Curve-based summary measures. *Stat Med* 1999; 18: 455-71.