

Curvas ROC

Autores: López de Ullibarri Galparsoro I, Píta Fernández, S.

Unidad de Epidemiología Clínica y Bioestadística. Complejo Hospitalario Juan Canalejo. A Coruña (España)

Cad Aten Primaria 1998; 5 (4): 229-235.

Actualización 25/09/2001.

Evaluación de pruebas diagnósticas

La toma de decisiones clínicas es un proceso extremadamente complejo en el que deberá finalmente ser valorada la utilidad para el manejo del paciente de cualquier prueba diagnóstica. En este contexto, es imprescindible conocer detalladamente la exactitud de las distintas pruebas diagnósticas, es decir, su capacidad para clasificar correctamente a los pacientes en categorías o estados en relación con la enfermedad (típicamente dos: estar o no estar enfermo, respuesta positiva o negativa a la terapia...).

Sensibilidad y Especificidad

Generalmente, la exactitud diagnóstica se expresa como sensibilidad y especificidad diagnósticas. Cuando se utiliza una prueba dicotómica (una cuyos resultados se puedan interpretar directamente como positivos o negativos), la **sensibilidad** es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada fracción de verdaderos positivos (FVP). La **especificidad** es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP).

Cuando los datos de una muestra de pacientes se clasifican en una tabla de contingencia por el resultado de la prueba y su estado respecto a la enfermedad, es fácil estimar a partir de ella la sensibilidad y la especificidad de la prueba ([tabla 1](#)). Conviene insistir –ya que esta distinción aparecerá repetidamente en lo sucesivo– en que lo que realmente obtenemos son *estimaciones* de los verdaderos valores de sensibilidad y especificidad para una población teórica de la que suponemos que nuestro grupo de pacientes constituye una muestra aleatoria. Por tanto, un tratamiento estadístico correcto de cantidades como las calculadas por el método descrito por la [tabla 1](#) exigirá incluir medidas de su precisión como estimadores, y, mejor aún, utilizarlas para construir intervalos de confianza para los verdaderos valores de sensibilidad y especificidad.

Tabla 1. Resultado de una prueba y su estado respecto a la enfermedad.			
		Verdadero Diagnóstico	
		Enfermo	Sano
Resultado de la Prueba	Prueba Positiva	Verdadero Positivo (VP)	Falso Positivo (FP)
	Prueba Negativa	Falso Negativo (FN)	Verdadero Negativo (VN)
		VP + FN	VN + FP
Sensibilidad	= $VP / (VP + FN)$ = FVP (fracción de verdaderos positivos)		
Especificidad	= $VN / (VN + FP)$ = FVN (fracción de verdaderos negativos) = $1 - FFP$ (fracción de falsos positivos)		

LA CURVA ROC

La limitación principal del enfoque hasta ahora expuesto estribaría en nuestra exigencia de que la respuesta proporcionada por la prueba diagnóstica sea de tipo dicotómico, por lo que en principio quedaría excluida la amplia gama de pruebas diagnósticas cuyos resultados se miden en una escala (nominalmente) continua o, al menos, discreta ordinal. Piénsese, por ejemplo, respecto al primer tipo en la determinación de la glucosa sérica por el laboratorio o, respecto al segundo, en una prueba realizada por el Servicio de Radiología en que los resultados se expresen empleando las categorías "seguramente normal", "probablemente normal", "dudoso", "probablemente anormal" y "seguramente anormal".

La generalización a estas situaciones se consigue mediante la elección de distintos *niveles de decisión* o *valores de corte* que permitan una clasificación dicotómica de los valores de la prueba según sean superiores o inferiores al valor elegido. La diferencia esencial con el caso más simple es que ahora contaremos no con un único par de valores de sensibilidad y especificidad que definan la exactitud de la prueba, sino más bien con un conjunto de pares correspondientes cada uno a un distinto nivel de decisión.

Este procedimiento constituye la esencia del análisis ROC, una metodología desarrollada en el seno de la Teoría de la Decisión en los años 50 y cuya primera aplicación fue motivada por problemas prácticos en la detección de señales por radar (aunque el detalle pueda parecer anecdótico, la equivalencia entre el operador que interpreta los picos en la pantalla del radar para decidir sobre la presencia de un misil y el médico que emplea el resultado de una prueba diagnóstica para decidir sobre la condición clínica del paciente, es completa ¹). La aparición del libro de Swets y Pickett ² marcó el comienzo de su difusión en el área de la Biomedicina, inicialmente en Radiología, donde la interpretación subjetiva de los resultados se recoge en una escala de clasificación, pero de modo creciente en relación con cualquier método diagnóstico que genere resultados numéricos.

Para centrar ideas, supongamos que, tanto para la población sana como para la enferma, la variable de decisión que representa el resultado de la prueba diagnóstica se distribuye normalmente, con media y desviación típica conocidas. En la [figura 1](#) se muestran las funciones de densidad de probabilidad para ambas variables, que mostrarán un determinado nivel de solapamiento. Si consideramos un valor arbitrario del resultado de la prueba, x –al que, en adelante, aludiremos como valor de corte–, la FVP (sensibilidad) y la FFP (1-especificidad) se corresponderán respectivamente con el área a la derecha de ese punto bajo la función de densidad de probabilidad de la población enferma (áreas clara y oscura) y de la población sana (área oscura). La curva ROC se obtiene representando, para cada posible elección de valor de corte, la FVP en ordenadas y la FFP en abscisas ([figura 2](#)).

Mediante esta representación de los pares (1-especificidad, sensibilidad) obtenidos al considerar todos los posibles valores de corte de la prueba, la curva ROC nos proporciona una representación global de la exactitud diagnóstica. La curva ROC es necesariamente creciente, propiedad que refleja el compromiso existente entre sensibilidad y especificidad: si se modifica el valor de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad. Si la prueba no permitiera discriminar entre grupos, la curva ROC sería la diagonal que une los vértices inferior izquierdo y superior derecho. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la discriminación fuera perfecta (100% de sensibilidad y 100% de especificidad) pasaría por dicho punto.

El modelo anterior, aplicable en principio a datos continuos, puede generalizarse al caso en que los datos se obtiene por algún sistema de clasificación en una escala discreta ordinal. Para ello basta suponer la existencia de unas variables latentes con distribución normal y de unos límites fijos que marcan los extremos de cada categoría. La [figura 3](#) muestra esquemáticamente este modelo para un ejemplo con cinco categorías.

Obviamente, el escenario en que hemos presentado la curva ROC es completamente teórico, por dos razones relacionadas entre sí:

- en la práctica no disponemos de las poblaciones (abstractas) de enfermos y sanos, sino simplemente de una muestra de ellas,

- en general, no conocemos las distribuciones de los valores de la prueba diagnóstica en dichas poblaciones.

Estas limitaciones nos obligan a considerar el problema práctico de la construcción de curvas ROC, que a continuación tratamos, desde un punto de vista típicamente estadístico.

MÉTODOS DE CÁLCULO DE LA CURVA ROC

Un primer grupo de métodos para construir la curva ROC lo constituyen los llamados métodos *no paramétricos*. Se caracterizan por no hacer ninguna suposición sobre la distribución de los resultados de la prueba diagnóstica. El más simple de estos métodos es el que suele conocerse como *empírico*, que consiste simplemente en representar todos los pares (FFP, FVP) – es decir todos los pares

(1-especificidad, sensibilidad) – para todos los posibles valores de corte que se puedan considerar con la muestra particular de que dispongamos. Desde un punto de vista técnico, este método sustituye las funciones de distribución teóricas por una estimación no paramétrica de ellas, a saber, la función de distribución empírica construida a partir de los datos. Informalmente, es como si en la [figura 1](#) sustituyéramos las funciones de densidad por histogramas obtenidos a partir de la muestra de pacientes sanos y enfermos y construyéramos la curva ROC a partir de ellos.

En la [figura 4](#) se representa la curva ROC obtenida por el método empírico para un conjunto de datos obtenidos en un grupo de pacientes investigados con el fin de establecer un diagnóstico de anemia ferropénica mediante la determinación del volumen corpuscular medio (ver apartado a) del apéndice). El verdadero diagnóstico se establece empleando como *gold standard* el examen de la médula ósea.

La representación obtenida por este método tiene forma aproximadamente en escalera. En efecto, para cada variación mínima del valor de corte que produzca cambios en sensibilidad o especificidad, al menos un caso pasa a ser considerado bien como verdadero positivo, lo que se corresponde con un trazo vertical, bien como falso positivo, lo que da lugar a un trazo horizontal. Existe aún otra posibilidad, derivada de la posibilidad de que se produzcan empates, es decir, dos o más casos con el mismo valor de la prueba: si el empate ocurre entre un caso del grupo enfermo y otro del grupo sano aparecerá un trazo diagonal en la representación.

Es evidente que este método es especialmente idóneo para datos de tipo continuo, sobre todo si la discretización (el redondeo) inducida por la precisión del método analítico utilizado no es muy importante, de modo que el número de empates sea proporcionalmente escaso. En este caso, la apariencia dentada de la curva es menos notoria a medida que crece el tamaño de la muestra e, idealmente, en el límite tendríamos una curva suave, la propia curva ROC teórica ([figura 1](#)). No obstante, también puede aplicarse a datos de tipo categórico. Claro está que ahora será inevitable la aparición de empates (al menos si el tamaño de la muestra es mayor que el número de categorías), con la consecuencia de que el gráfico consistirá, independientemente del tamaño de la muestra, en un número fijo de líneas en general diagonales que unen los puntos correspondientes a los pares (1-especificidad, sensibilidad) calculados para cada categoría. En la [figura 5](#) se presenta un ejemplo de la aplicación de este método a un conjunto de datos procedente de la clasificación en cinco categorías de imágenes obtenidas por tomografía computerizada (ver [apartado b\) del apéndice](#)).

Una práctica frecuente que desaconsejamos es realizar la agrupación artificial de los resultados de una prueba diagnóstica con valores continuos en un número limitado de categorías, como paso previo a la construcción de la curva ROC. Si no se dispone de medios de representación automáticos este recurso permite simplificar los cálculos necesarios para construir la curva ROC, pero entre otras deficiencias, es evidente que así se obtiene una estimación sesgada de la verdadera curva ROC, concretamente una infravaloración de la verdadera exactitud diagnóstica ^{3,4}.

Existen otros métodos no paramétricos ⁵ aplicables a datos continuos que permiten obtener curvas ROC suavizadas, en contraposición con la forma dentada de la curva obtenida por el método empírico. La idea es básicamente obtener estimaciones no paramétricas suavizadas de las funciones de densidad de las dos distribuciones de resultados de la prueba empleando generalmente estimadores de tipo núcleo. A partir de

dichas densidades –en lugar de a partir de los histogramas, como en el método anterior– se obtiene directamente la curva ROC que, como dijimos, será suave.

Los métodos *paramétricos* se basan en postular un determinado tipo de distribución para la variable de decisión en las dos poblaciones que se trata de distinguir². El modelo más frecuentemente utilizado es el binormal, que supone la normalidad de las variables tanto en la población sana como en la enferma, pero existen muchos otros modelos posibles que surgen al considerar distintas distribuciones, similares a la normal como la logística (modelo biológico) o no, como la exponencial negativa. El problema ahora se reduce a estimar los parámetros de cada distribución por un método estadísticamente adecuado, en general el método de máxima verosimilitud. Se obtiene así una curva ROC suave, pero puede ocurrir una sustancial falta de ajuste si los supuestos distribucionales resultan ser erróneos^{3,4}. Por ello, si se va a emplear este método debe previamente someterse la hipótesis sobre la naturaleza de las distribuciones a un contraste de significación. También es recomendable emplear una transformación de la variable inicial que logre que los datos sean más compatibles con las distribuciones asumidas, aunque este juicio sólo pueda basarse en un examen en gran medida visual y por lo tanto esté expuesto a interpretaciones subjetivas. Estas limitaciones hacen que el método no sea en general recomendable para datos continuos. Su utilidad es mayor con datos discretos: varios investigadores han examinado el modelo binormal para datos de clasificación, sin encontrar situaciones en las que el modelo fallara seriamente^{6,7}. De hecho, esta última observación constituye la base para un método aplicable tanto a datos continuos como de clasificación, debido a Metz et al.⁸ (el software que lo implementa está libremente disponible en Internet en <http://www-radiology.uchicago.edu/krl/toppage11.htm>). Según este método, primero se agrupan los datos en categorías ordenadas y después se aplica un algoritmo paramétrico para crear una curva ROC suave. Del método se dice que es *semiparamétrico*^{5,9}, porque aunque supone la existencia de una transformación que haga que las dos distribuciones sean aproximadamente normales, ésta se deja sin especificar. La dependencia mucho menor de la validez de las asunciones se debe principalmente a la invariancia de la curva ROC frente a las transformaciones monótonas de la escala de la variable de decisión⁸. Investigaciones recientes parecen indicar que el método se comporta empíricamente bien en una amplia variedad de situaciones¹⁰. Permanece, no obstante, el problema de que el ajuste no es reproducible a menos que el esquema de categorización empleado sea objetivo y esté estandarizado. Las [figuras 4](#) y [5](#) muestran las curvas ROC ajustadas por este método y permiten compararlas a las obtenidas por el método empírico.

ANÁLISIS ESTADÍSTICO DE LAS CURVAS ROC

Área bajo la curva

Como observamos más arriba, la mayor exactitud diagnóstica de una prueba se traduce en un desplazamiento "hacia arriba y a la izquierda" de la curva ROC. Esto sugiere que el área bajo la curva ROC (ABC) se puede emplear como un índice conveniente de la exactitud global de la prueba: la exactitud máxima correspondería a un valor de ABC de 1 y la mínima a uno de 0.5 (si fuera menor de 0.5 debería invertirse el criterio de positividad de la prueba).

En términos probabilísticos, si X_E y X_S son las dos variables aleatorias que representan los valores de la prueba en las poblaciones enferma y sana, respectivamente, puede probarse que el ABC de la "verdadera" curva ROC (intuitivamente, aquella que obtendríamos si el tamaño de la muestra fuera infinito y la escala de medida continua) es precisamente $\theta = \Pr\{X_E > X_S\}$, o, en palabras, la probabilidad de que, si se eligen al azar un paciente enfermo y otro sano, sea mayor el valor de la prueba en aquél que en éste¹¹.

Cuando la curva ROC se genera por el método empírico, independientemente de que haya empates o no, el área puede calcularse mediante la regla trapezoidal, es decir, como la suma de las áreas de todos los rectángulos y trapecios (correspondientes a los empates) que se pueden formar bajo la curva. Estadísticamente, la observación importante, puesto que permite hacer contrastes de significación y dar intervalos de confianza para la verdadera área bajo la curva, es que el área calculada por el método geométrico anterior coincide con el valor del estadístico de suma de rangos de Wilcoxon, W ¹². Esto no es sorprendente, ya que, cuando X_A y X_B son dos variables aleatorias independientes cualesquiera, dicho estadístico es conocido precisamente por su uso para contrastar la hipótesis $\Pr\{X_A > X_B\} = 1/2$, que en nuestro contexto es la hipótesis nula de que ABC sea $1/2$, es decir, de que la prueba no sea capaz de discriminar entre los dos grupos. Hanley y McNeil¹¹ dan fórmulas tanto para el estadístico W como para

su error estándar y discuten el problema de la estimación de este último. En general, se suelen dar intervalos de confianza del ABC contruidos de la manera estándar, v. g. al nivel de confianza del 95% intervalos de extremos $W \pm 1.96*EE(W)$, siendo $EE(W)$ una estimación del error estándar de W . Con los datos del volumen corpuscular medio, el ABC calculado por este método es 0.699 y un intervalo de confianza del 95% (IC 95%) es (0.585,0.813).

Cuando se ajusta un modelo como el binormal empleando técnicas estadísticas se obtienen, además de estimaciones de los parámetros que definen la curva ROC, estimaciones del ABC y de su error estándar, que pueden emplearse para construir intervalos de confianza y efectuar contrastes de significación como en el caso no paramétrico. Empleando el método semiparamétrico de Metz, el ABC estimado con los datos de volumen corpuscular medio es 0.703 y (0.591,0.799) un IC 95%.

Cuando el número de empates es elevado, como ocurre cuando se emplean datos de clasificación, el estadístico W –el área calculada por el método empírico– proporciona un estimador sesgado de la verdadera área, lo que hace recomendable emplear un método distinto, por ejemplo uno basado en un método paramétrico. Con los datos de la tomografía computerizada el ABC estimado es 0.893 (IC 95% (0.830,0.956)) ó 0.911 (IC 95% (0.838,0.956)) según se emplee el estadístico de Wilcoxon o el método de Metz en su cálculo, respectivamente, lo que sugiere que en este caso la magnitud del sesgo no es importante.

Área parcial

Son imaginables situaciones en las que las propias características ventajosas del ABC se conviertan en un inconveniente para su uso clínico. El ABC puede interpretarse como un promedio de la sensibilidad (especificidad) sobre todos los valores posibles de especificidad (sensibilidad). Puede que clínicamente sólo nos interesen los puntos de la curva ROC que aseguren altos valores de sensibilidad o especificidad. Un caso típico es el de las mamografías en programas de detección precoz del cáncer, donde debe asegurarse una alta sensibilidad de la prueba¹³. Se han propuesto índices de área parcial que pueden ser empleados para evaluar la exactitud restringida a los puntos de operación de interés de la curva ROC^{13,14}.

Comparación de dos pruebas

Cuando se dispone de dos (o más) pruebas para abordar el diagnóstico de un mismo problema clínico, el cálculo del ABC brinda un método conveniente para comparar globalmente su exactitud diagnóstica relativa. En principio, al comparar dos pruebas preferiremos la que tenga mayor ABC, por ser la de mayor exactitud diagnóstica de las dos. Desde un punto de vista estadístico el problema es valorar si la diferencia observada entre las ABC calculadas para dos pruebas distintas es debida a la variabilidad inherente al muestreo o es más bien atribuible a una diferencia real en la exactitud de ambas pruebas. Podemos enunciar este problema como un contraste de la hipótesis nula de igualdad de las dos ABC, que denotaremos por ABC_A y ABC_B , frente a una alternativa bilateral. En general, se dispone de los valores para las dos pruebas en una única muestra de pacientes. El contraste debido a Hanley y McNeil¹⁵, que podemos considerar representativo de los desarrollados en esta situación, utiliza como estadístico del contraste

$$z = (ABC_A - ABC_B) / \sqrt{EE_A^2 + EE_B^2 - 2rEE_AEE_B}$$

siendo ABC el área observada, EE el error estándar del ABC y r la correlación entre ABC_A y ABC_B . Al nivel de significación α se rechaza la hipótesis nula cuando $|z| > z_{\alpha/2}$, siendo $z_{\alpha/2}$ el cuantil de orden $1-\alpha$

/2 de una distribución normal estándar, v. g. si $\alpha = 0.05$ es $z_{\alpha/2} = 1.96$.

No obstante, hagamos notar que la comparación entre dos pruebas no debe reposar exclusivamente en contrastes como el anterior. Pueden existir dos pruebas con sendas curvas ROC muy distintas de forma, hecho que puede tener importantes implicaciones prácticas, y que, sin embargo, sean prácticamente iguales respecto a su ABC (figura 6). El empleo del área parcial puede permitir manejar correctamente estas situaciones. En cualquier caso, es evidente que nunca debería prescindirse de un examen visual detenido de un gráfico que muestre simultáneamente ambas curvas ROC.

Elección del valor de corte

El empleo en la práctica médica de una prueba diagnóstica exige la elección de un valor de corte. Para ello es imprescindible un conocimiento detallado de los riesgos y beneficios de las decisiones médicas derivadas del resultado de la prueba. Un enfoque sencillo ^{3,16}, que utiliza la razón de costes de un resultado falso positivo frente a un falso negativo, requiere calcular el coeficiente

$$m = \frac{\text{coste de falsos positivos}}{\text{coste de falsos negativos}} * \frac{1 - P}{P}$$

donde P es la prevalencia de la enfermedad. El valor de corte óptimo se determina hallando el punto de la curva ROC (que supondremos suave) con la siguiente propiedad: la tangente a la curva en ese punto tiene pendiente m.

Incluso una fórmula sencilla como la anterior deja en evidencia que en la mayoría de los casos nuestra pretensión de calcular un valor de corte óptimo será excesiva, salvo que uno se contente con estimaciones imprecisas o puramente intuitivas. De hecho, es un problema que se aborde más adecuadamente con otras herramientas más complejas disponibles en el seno del Análisis de Decisiones Clínicas ^{17,18}.

Apéndice

Apéndice 1.a. Volumen corpuscular medio eritrocitario en 100 pacientes con posible diagnóstico de anemia ferropénica. El verdadero diagnóstico se establece por biopsia de la médula ósea.	
Examen de la médula ósea	Volumen corpuscular medio
Ausencia de hierro ($n_E = 33$)	52, 58, 62, 65, 67, 68, 70, 72, 72, 73, 73, 74, 75, 77, 77, 77, 78, 78, 80, 80, 81, 81, 81, 81, 84, 84, 85, 85, 87, 88, 88, 92, 92
Presencia de hierro ($n_S = 67$)	60, 66, 68, 69, 71, 71, 71, 73, 74, 74, 74, 76, 76, 77, 77, 77, 78, 79, 79, 79, 80, 80, 81, 81, 82, 82, 82, 83, 83, 83, 83, 83, 83, 83, 84, 84, 84, 85, 85, 86, 86, 86, 86, 88, 88, 88, 89, 89, 89, 90, 90, 90, 91, 91, 93, 93, 93, 94, 94, 94, 94, 96, 97, 98, 100, 103
Modificado de Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. <i>Arch Pathol Lab Med</i> 1986; 110: 13-20.	

Apéndice 1.b. Clasificación de 109 imágenes de tomografía computerizada.					
Verdadera condición	Seguramente normal	Probablemente normal	Dudosa	Probablemente anormal	Seguramente anormal
Normal	33	6	6	11	2
Anormal	3	2	2	11	33
Tomado de Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. <i>Radiology</i> 1982; 143: 29-36.					

Figura 1. Distribución de los resultados de una prueba en las poblaciones de pacientes sanos y enfermos. Las distribuciones teóricas se suponen normales de media y desviación típica diferentes. Las áreas a la derecha del valor de corte x corresponden a la sensibilidad (áreas clara y oscura) y 1-especificidad (área oscura).

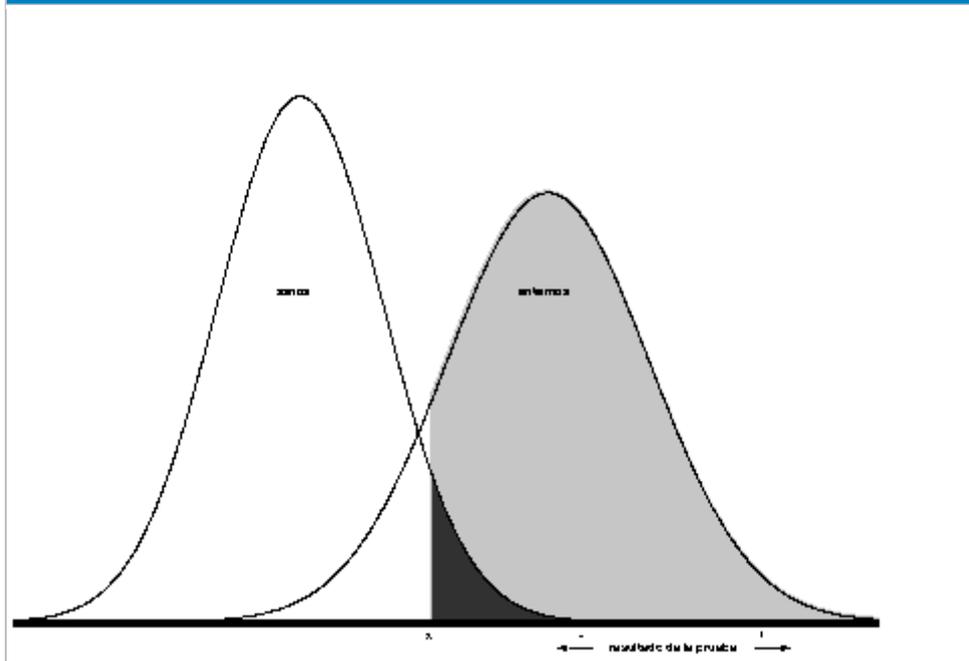


Figura 2. Curva ROC correspondiente a la distribución teórica de los resultados de una prueba representada en la figura 1. Se muestra el punto correspondiente al valor de corte x .

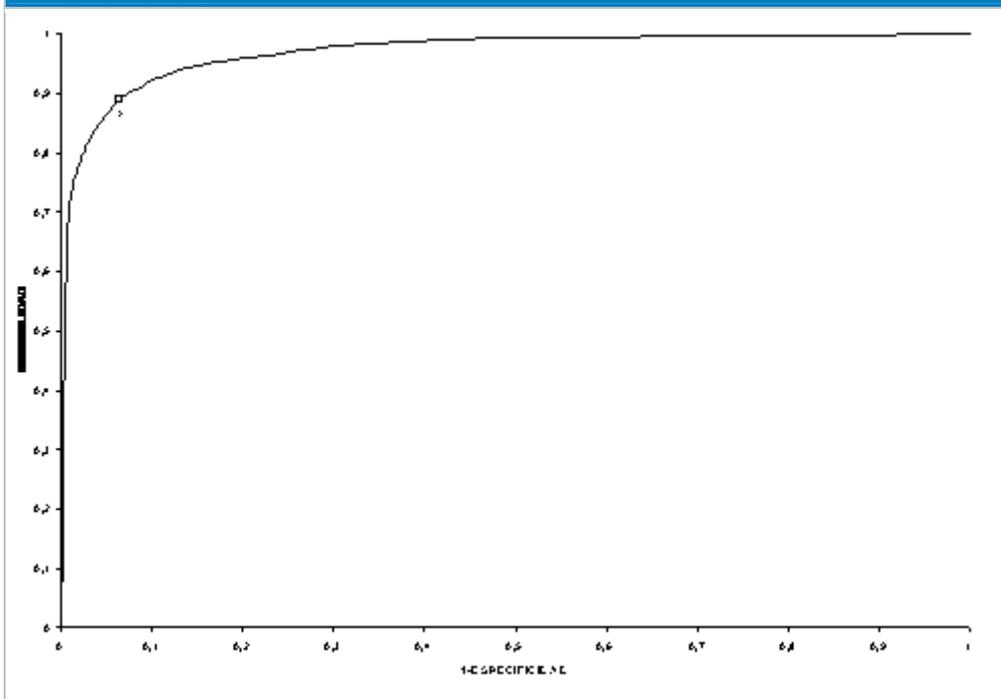


Figura 3. Representación esquemática de un modelo para datos de clasificación con cinco categorías.

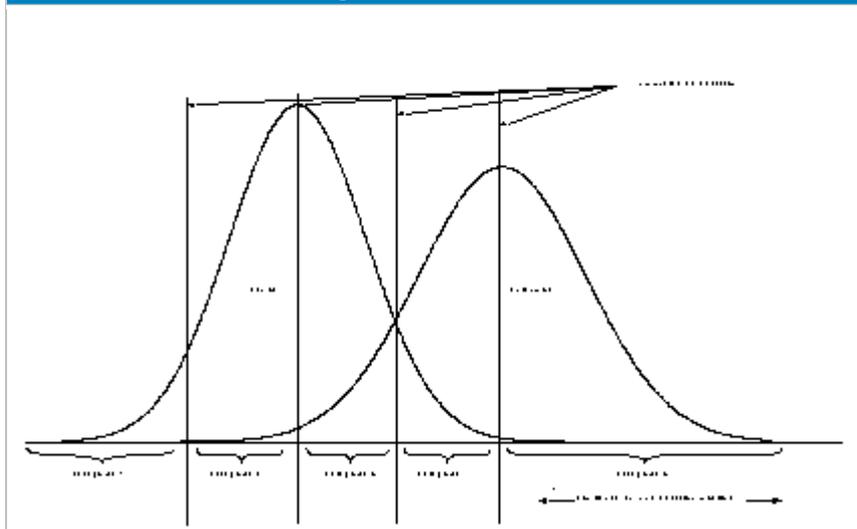


Figura 4. Curvas ROC calculadas por los métodos empírico y semiparamétrico para los datos de volumen corpuscular medio.

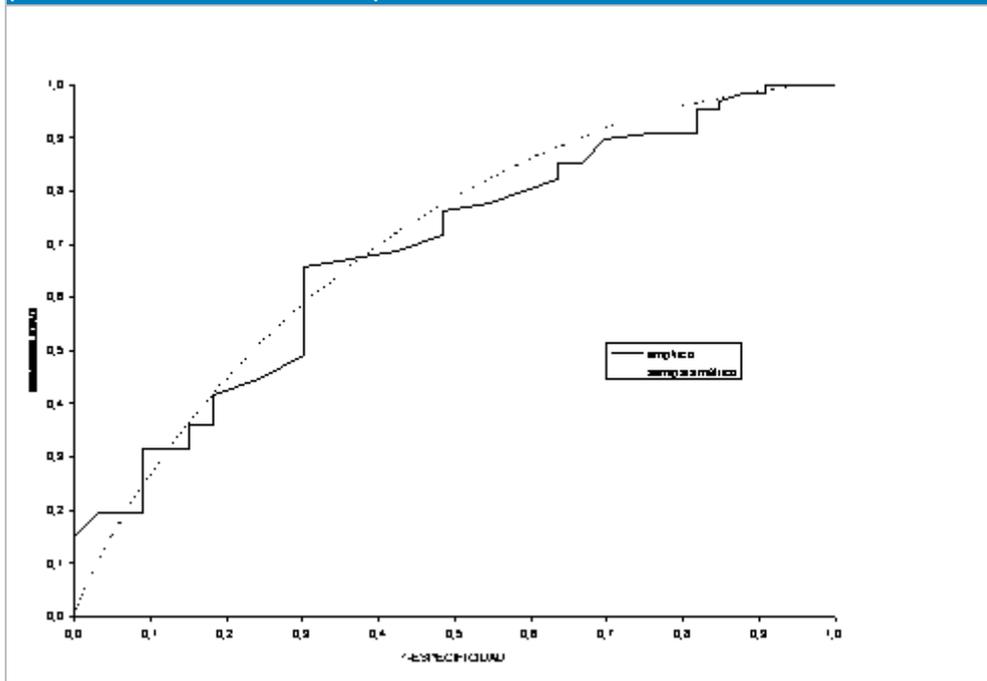


Figura 5. Curvas ROC calculadas por los métodos empírico y semiparamétrico para los datos de tomografía computerizada.

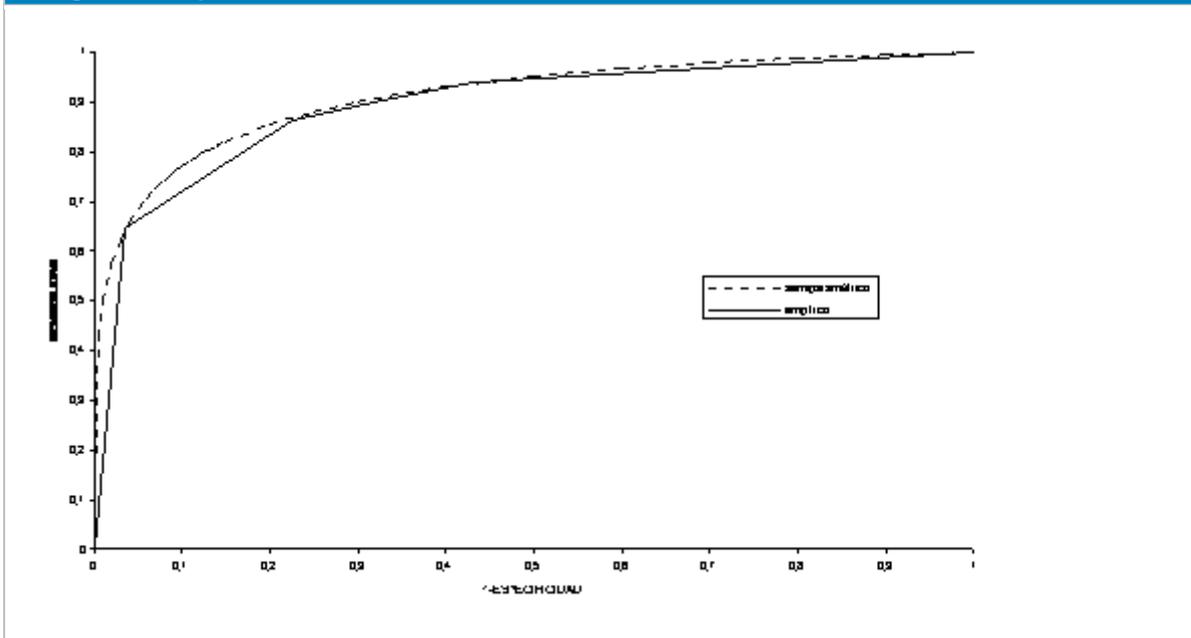
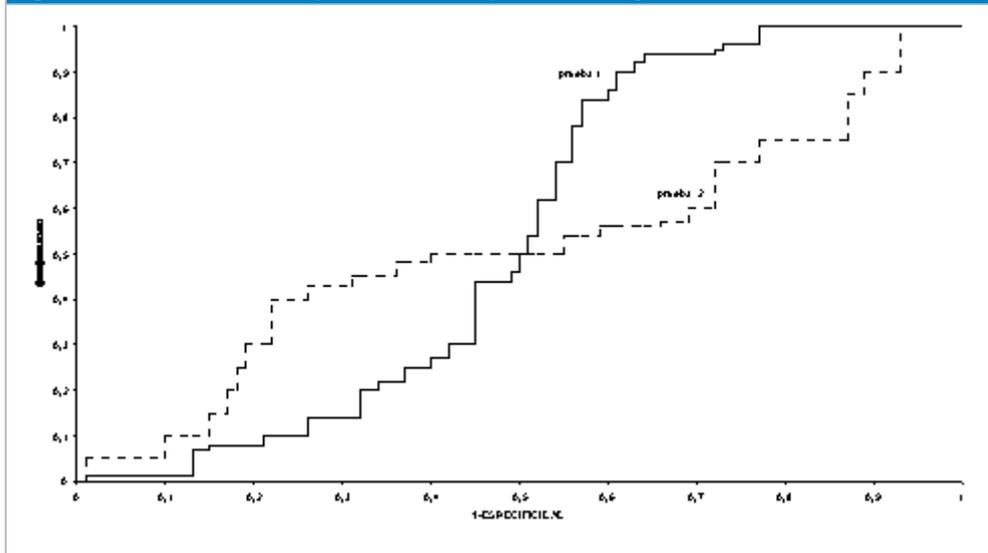


Figura 6. Curvas ROC empíricas de dos pruebas diagnósticas distintas.



Bibliografía

1. Robertson EA, Zweig MH. Use of receiver operating characteristic curves to evaluate the clinical performance of analytical systems. *Clin Chem* 1981; 27: 1569-1574. [[Medline](#)]
2. Swets JA, Pickett RM. *Evaluation of diagnostic systems: methods from signal detection theory*. Nueva York: Academic Press; 1982.
3. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39: 561-577. [[Medline](#)]
4. Burgueño MJ, García-Bastos JL, González-Buitrago JM. Las curvas ROC en la evaluación de las pruebas diagnósticas. *Med Clin (Barc)* 1995; 104: 661-670. [[Medline](#)]
5. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statist Med* 1997; 16: 2143-2156. [[Medline](#)]
6. Hanley JA. The robustness of the binormal model used to fit ROC curves. *Med Decision Making* 1988; 8: 197-203. [[Medline](#)]
7. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psych Bull* 1986; 99: 181-198. [[Medline](#)]
8. Metz CE, Herman BA, Shen, J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statist Med* 1998; 17: 1033-1053. [[Medline](#)]
9. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann Statist* 1996; 24: 25-40.
10. Hanley JA. The use of the "binormal" model for parametric ROC analysis of quantitative diagnostics tests. *Statist Med* 1996; 15: 1575-1585. [[Medline](#)]
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36. [[Medline](#)]
12. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *J Math Psych* 1975; 12: 387-415.
13. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostics tests. *Radiology* 1996; 201: 745-750.
14. McClish DK. Analyzing a portion of the ROC curve. *Med Decision Making* 1989; 9: 190-195. [[Medline](#)]
15. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148: 839-843. [[Medline](#)]
16. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. *N Engl J Med* 1975; 293: 211-215. [[Medline](#)]
17. Weinstein MC, Fineberg HV. *Clinical Decision Analysis*. Philadelphia: WB Saunders Co; 1980.
18. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch Pathol Lab Med* 1986; 110: 13-20. [[Medline](#)]