Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos

Jaime Cerda y Lorena Cifuentes

Pontificia Universidad Católica de Chile, Santiago.
Facultad de Medicina
Departamento de Salud Pública. (JC).
División de Pediatría (LC).
Programa de Salud Basada en Evidencia (JC, LC).

Los autores reportan no tener conflictos de interés

Recibido: 12 de julio de 2011
Aceptado: 25 de julio de 2011

Correspondencia a:
Jaime Cerda Lorca
jcerda@med.puc.cl

Using ROC curves in clinical investigation. Theoretical and practical issues

Numerous diagnostic tests report their results quantitatively, using continuous scales. *Receiver operating characteristic curve* (ROC) analysis provides a statistical method for the assessment of the diagnostic accuracy of these tests, being used for three specific purposes: determine of the cutoff value with the highest sensitivity and specificity, evaluate the discriminative capacity of the diagnostic test, in other words, its ability to differentiate healthy versus sick individuals, and compare the discriminative capacity of two or more diagnostic tests that express their results as continuous scales. Based on a real clinical investigation, this article illustrates theoretical aspects regarding the construction of ROC curves, being its objective to help readers and investigators interpret correctly their results.

Key words: Diagnostic tests, ROC curve, biostatistics.
Palabras clave: Tests diagnósticos, curva ROC, bioestadística.

Introducción

En dos artículos anteriores publicados en esta revista fueron analizadas las propiedades de un test diagnóstico[1] y su aplicación clínica[2]. Sin embargo, la teoría expuesta y los ejemplos citados en ellos incluyeron únicamente estudios con resultados categóricos, sean estos dicotómicos (e.g. positivo, negativo) o politómicos (e.g. leve, moderado y severo). La realidad nos indica que una amplia gama de tests diagnósticos reportan sus resultados cuantitativamente, utilizando escalas continuas (e.g. recuento de leucocitos, proteína C reactiva). En ellas, para formular el diagnóstico de una determinada enfermedad se establece un punto de corte, sobre el cual se apoya la presencia del diagnóstico y bajo el cual se rechaza, o viceversa. El análisis en base a curvas ROC *(receiver operating characteristic curve)* constituye un método estadístico para determinar la exactitud diagnóstica de tests que utilizan escalas continuas, siendo utilizadas con tres propósitos específicos: determinar el punto de corte en el que se alcanza la sensibilidad y especificidad más alta, evaluar la capacidad discriminativa del test diagnóstico, es decir, su capacidad de diferenciar sujetos sanos *versus* enfermos, y comparar la capacidad discriminativa de dos o más tests diagnósticos que expresan sus resultados como escalas continuas[3,4]. En base a una investigación real, el presente artículo ilustra el sustrato teórico subyacente a la construcción de curvas ROC, siendo su objetivo ayudar a lectores e investigadores a interpretar correctamente sus resultados.

¿Qué es una curva ROC?

Para explicar qué es una curva ROC utilizaremos como ejemplo la concentración de procalcitonina sérica (PCT), postulada como test diagnóstico para endocarditis infecciosa en sujetos con sospecha de presentar dicha patología. La concentración de PCT sérica corresponde a una escala continua, cuya mediana es significativamente diferente en los pacientes con y sin endocarditis infecciosa (6,56 *vs* 0,44 ng/ml, respectivamente)[5]. Operativamente, para poder diagnosticar una endocarditis infecciosa a partir de un valor de PCT sérica es necesario establecer un punto de corte, considerándose "positivos" aquellos valores mayores o iguales al punto de corte, y "negativos" aquellos valores menores al punto de corte. Al contrastar los resultados de la PCT sérica con los resultados de un estándar de oro para el diagnóstico de endocarditis infecciosa (i.e. criterios de Duke) surgen cuatro posibles alternativas diagnósticas:

1. Pacientes que tienen una endocarditis infecciosa según el estándar de oro, y cuya medición de PCT sérica los diagnostica como "positivos". Estos pacientes reciben el nombre de "verdaderos positivos, VP".
2. Pacientes que tienen una endocarditis infecciosa según el estándar de oro, sin embargo, su medición de PCT sérica los diagnostica como "negativos". Estos pacientes reciben el nombre de "falsos negativos, FN".
3. Pacientes que no tienen una endocarditis infecciosa según el estándar de oro, y cuya medición de PCT sérica los diagnostica como "negativos". Estos pacientes reciben el nombre de "verdaderos negativos, VN".

4. Pacientes que no tienen una endocarditis infecciosa según el estándar de oro, sin embargo, su medición de PCT sérica los diagnostica como "positivos". Estos pacientes reciben el nombre de "falsos positivos, FP".

Conociendo el número de sujetos de cada una de estas alternativas diagnósticas es posible calcular la sensibilidad y especificidad para el punto de corte que les origina, según las siguientes fórmulas: sensibilidad = VP/(VP+FN) y especificidad = VN/(FP+VN) (Tabla 1). El lector notará que existen tantos puntos de corte posibles de PCT sérica como valores posee su escala de medición, y que cada punto de corte posee su respectiva sensibilidad y especificidad para el diagnóstico de endocarditis infecciosa. Un gráfico de curva ROC ilustra la sensibilidad y especificidad de cada uno de los posibles puntos de corte de un test diagnóstico cuya escala de medición es continua. La curva ROC se construye en base a la unión de distintos puntos de corte, correspondiendo el eje Y a la sensibilidad y el eje X a (1-especificidad) de cada uno de ellos. Ambos ejes incluyen valores entre 0 y 1 (0% a 100%). A modo de referencia, en todo gráfico de curva ROC se traza una línea desde el punto 0,0 al punto 1,1, llamada diagonal de referencia o línea de no-discriminación (concepto a abordar más adelante). La Figura 1 ilustra el gráfico de curva ROC de un test diagnóstico hipotético.

Tabla 1. Categorías diagnósticas obtenidas a partir de un estudio de exactitud diagnóstica

	Estándar de oro Positivo	Estándar de oro Negativo
Test positivo	Verdaderos positivos (VP)	Falsos positivos (FP)
Test negativo	Falsos negativos (FN)	Verdaderos negativos (VN)

Sensibilidad = VP/(VP+FN), Especificidad = VN/(FP+VN).

¿Qué punto de corte de la escala continua determina la sensibilidad y especificidad más alta?

Los estudios que evalúan la exactitud diagnóstica de un test siguen en su mayoría un diseño transversal, en el cual un mismo sujeto es evaluado concomitantemente mediante el test diagnóstico y un estándar de oro. En el ejemplo[5], los investigadores enrolaron 67 pacientes consecutivos con sospecha de endocarditis infecciosa, a quienes midieron su concentración de PCT sérica y aplicaron el estándar de oro, consistente en los criterios de Duke[6]. Conocida la condición de "enfermo" y "no-enfermo" a partir del estándar de oro, clasificaron a los pacientes en VP, VN, FP y FN, y procedieron a calcular la sensibilidad y especificidad para cada concentración de PCT sérica (es decir, para cada posible punto de corte de la escala continua). En base a dichos cálculos construyeron una curva ROC, viéndose enfrentados a un primer desafío: identificar el punto de corte de PCT sérica que determina la sensibilidad y especificidad más alta.

El punto de corte de una escala continua que determina la sensibilidad y especificidad más alta es aquel que presenta el mayor índice de Youden, calculado según la fórmula (sensibilidad + especificidad − 1). Gráficamente, éste corresponde al punto de la curva ROC más cercano al ángulo superior−izquierdo del gráfico (punto 0,1), es decir, más cercano al punto del gráfico cuya sensibilidad = 100% y especificidad = 100% (Figura 2). En este momento es preciso hacer una aclaración: el índice de Youden identifica el punto de corte que determina la sensibilidad y especificidad más alta *conjuntamente* (i.e. para un mismo punto), sin embargo, dicho punto de corte no necesariamente determina la sensibilidad ni la especificidad más alta que podría alcanzar el test (generalmente la sensibilidad más alta es determinada por un punto de corte, mientras que la especificidad más alta es determinada por otro).

Existen situaciones en las que se requiere disponer de un test diagnóstico altamente sensible (e.g. tamizaje de enfermedades) o bien altamente específico (e.g. confirmación de enfermedades). En tales circunstancias, no es aconsejable utilizar el punto de corte identificado por el índice de Youden; por el contrario, resulta más útil conocer los valores de sensibilidad y especificidad determinados por diferentes puntos de corte, y optar por aquel que determine la mayor sensibilidad, o la mayor especificidad, según sea el objetivo.

En el ejemplo[5], el punto de corte de concentración de PCT sérica ubicado en 2,3 ng/ml determinó la sensibilidad y especificidad más alta *conjuntamente* para el diagnóstico de endocarditis infecciosa, siendo su sensibilidad = 81% y especificidad = 85%. Su índice de Youden equivale a

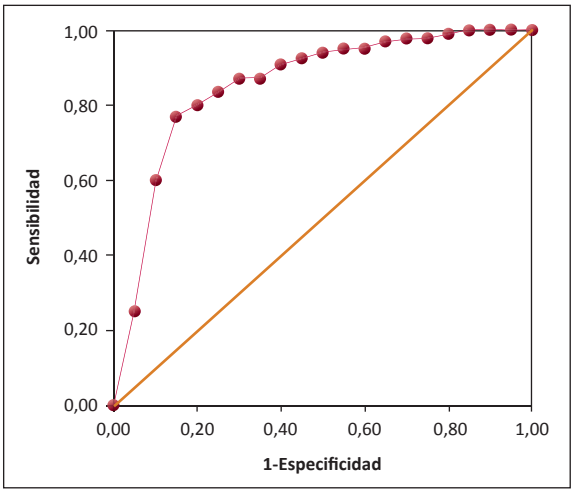

Figura 1. Gráfico de curva ROC de un test diagnóstico hipotético. Cada punto de la curva ROC (círculos negros) corresponde a un posible punto de corte del test diagnóstico, y nos informa su respectiva sensibilidad (eje Y) y 1-especificidad (eje X). Ambos ejes del gráfico incluyen valores entre 0 y 1 (0% a 100%). La línea trazada desde el punto 0,0 al punto 1,1 recibe el nombre de diagonal de referencia, o línea de no-discriminación.

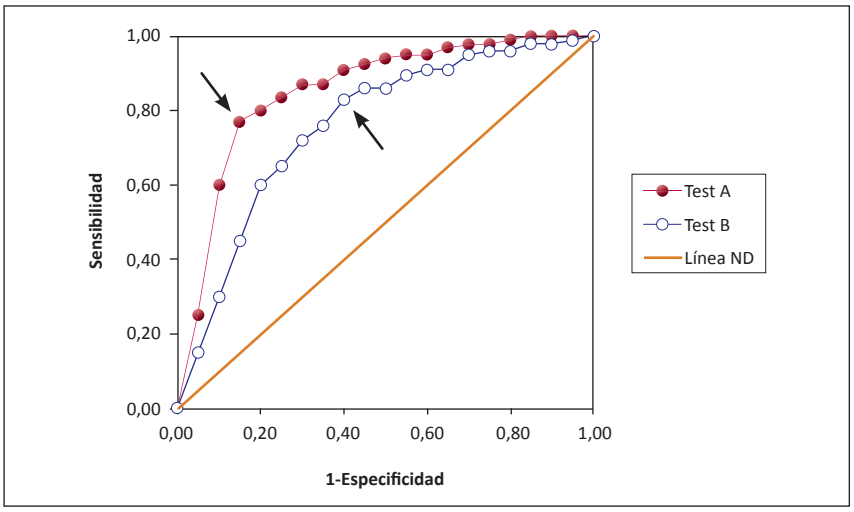

Figura 2. Gráfico de curva ROC de dos tests diagnósticos hipotéticos (A y B), y línea de no-discriminación (línea ND). Para cada curva ROC, las flechas indican el punto de corte que determina la sensibilidad y especificidad conjuntas más alta.

(0,81 + 0,85 – 1) = 0,66.

Evaluación de la capacidad discriminativa de un test diagnóstico

La capacidad discriminativa de un test diagnóstico se refiere a su habilidad para distinguir pacientes sanos *versus* enfermos. Para ello, el parámetro a estimar es el área bajo la curva ROC (AUC, *area under the curve*), medida única e independiente de la prevalencia de la enfermedad en estudio. El AUC refleja qué tan bueno es el test para discriminar pacientes con y sin la enfermedad a lo largo de todo el rango de puntos de corte posibles.

Para comprender de mejor manera el concepto de discriminación, es más simple pensar que el eje Y del gráfico de curva ROC corresponde a la proporción de verdaderos positivos sobre el total de pacientes enfermos (i.e. sensibilidad), y que el eje X corresponde a la proporción de falsos positivos sobre el total de sujetos sanos (i.e. 1–especificidad). Visto de esta manera, un gráfico de curva ROC ilustra la "proporción de verdaderos positivos" (eje Y) *versus* la "proporción de falsos positivos" (eje X) para cada punto de corte de un test diagnóstico cuya escala de medición es continua. Como fue mencionado anteriormente, a modo de referencia se traza una línea desde el punto 0,0 al punto 1,1 (diagonal de referencia, o línea de no-discriminación). Este línea describe lo que sería la curva ROC de un test diagnóstico incapaz de discriminar pacientes sanos *versus* enfermos, debido a que cada punto de corte que la compone determina la misma proporción de verdaderos positivos y de falsos positivos. Un test diagnóstico tendrá mayor capacidad discriminativa en la medida que sus puntos de corte tracen una curva ROC lo más lejana posible a la línea de no-discriminación; dicho de otra manera, lo más cercana posible a los lados izquierdo y superior del gráfico.

Mencionamos anteriormente que los ejes del gráfico de curva ROC adoptan valores entre 0 y 1 (0% y 100%), delimitando un cuadrado de área = 1,00. Un test diagnóstico se considera no-discriminativo si su curva ROC coincide con la línea de no-discriminación, la cual posee AUC = 0,50 (notará el lector que la línea de no-discriminación divide en dos mitades iguales el cuadrado de área = 1,00, razón por la cual decimos que su AUC = 0,50). A medida que el AUC de un test diagnóstico se acerca al valor 1,00 (test diagnóstico perfecto), mayor será su capacidad discriminativa.

En el ejemplo, el AUC de la concentración de PCT sérica fue 0,86 y su intervalo de confianza 95% (IC 95%) fue 0,75-0,96. No existe un valor de AUC a partir del cual se considere que un test diagnóstico es capaz de discriminar pacientes sanos *versus* enfermos. Sin embargo, si consideramos que un AUC = 0,75 se encuentra a medio camino entre la no-discriminación (AUC = 0,50) y la discriminación perfecta (AUC = 1,00), el AUC de la concentración de PCT sérica se encuentra más cercana a la perfección que a la no-discriminación, por lo tanto, resulta razonable plantear que la PCT sérica es un test diagnóstico con una capacidad aceptable de discriminar pacientes con y sin endocarditis infecciosa. Es importante mencionar que el AUC es un estimador muestral de un parámetro poblacional, por ello los investigadores reportaron su IC 95%. Si este intervalo incluyese el valor 0,50 (p. ej.; IC 95% 0,45-0,96), no sería posible afirmar que el AUC de la concentración de PCT es diferente a la no-discriminación.

Comparación de la capacidad discriminativa de dos tests diagnósticos

Para comparar la capacidad discriminativa de dos tests diagnósticos es importante verificar un concepto metodológico de suma importancia: los tests a comparar deben ser medidos simultáneamente, aplicados sobre los mismos sujetos y contrastados contra el mismo estándar de oro. Verificados estos requisitos, para comparar la capacidad discriminativa de dos tests diagnósticos deben compararse sus respectivas AUC, siendo más discriminativo el test con la mayor AUC. En el ejemplo[5], se comparó la capacidad discriminativa de la concentración de PCT sérica con la concentración de proteína C-reactiva (PCR), siendo el AUC de la primera 0,86 (IC95% 0,75-0,96) y de la segunda 0,66 (IC95% 0,51-0,80). Estos valores sugieren que la PCT sérica discrimina de mejor manera que la PCR los pacientes con y sin endocarditis infecciosa, debido a que el AUC de la PCT sérica fue mayor que el AUC de la PCR. Sin embargo, para poder afirmar (y no sugerir) que existe una diferencia significativa entre el

AUC de la PCT sérica y de la PCR es necesario comparar estadísticamente ambas áreas bajo la curva ROC, según los métodos descritos por Hanley & McNeil[7] o DeLong[8] (este último de preferencia). Por desgracia estos tests no están incluidos en la totalidad de softwares estadísticos que permiten construir curvas ROC, siendo una práctica no infrecuente por parte de algunos investigadores afirmar que un test diagnóstico "X" tiene mayor capacidad discriminativa que un test diagnóstico "Y" solamente porque el AUC de "X" es numéricamente mayor al AUC de "Y" (esta práctica solamente permite *sugerir* que "X" tiene mayor capacidad discriminativa de "Y", mas no permite afirmarlo).

En el ejemplo[5] solamente es posible sugerir que PCT sérica tiene una mayor capacidad discriminativa que PCR para el diagnóstico de endocarditis infecciosa, debido a que los investigadores no reportaron haber realizado una comparación estadística de sus respectivas AUC.

Consideraciones finales

Para una construcción e interpretación correcta de gráficos de curvas ROC es necesario tener en cuenta tres conceptos finales. En primer lugar, la sensibilidad, especificidad y AUC son estimadores muestrales de parámetros poblacionales; por consiguiente, cada uno tiene asociado un error de estimación, siendo necesario reportar sus respectivos intervalos de confianza. En segundo lugar, los estudios de exactitud diagnóstica a partir de los cuales se construyen curvas ROC corresponden generalmente a diseños de tipo transversal. La validez de estos estudios (i.e. riesgo de sesgo) debe ser evaluada críticamente, siguiendo las recomendaciones descritas por la Medicina Basada en Evidencia[9]. Al respecto, un análisis de curva ROC estadísticamente perfecto carece de todo sentido si los datos utilizados para construir la curva ROC provienen de un estudio metodológicamente deficiente (desde un punto de vista metodológico, el estudio de Mueller y cols[5], corresponde a uno de exactitud diagnóstica, cumpliéndose los siguientes indicadores de calidad: espectro de población adecuado, estándar de oro adecuado, interpretación ciega del test y del estándar de oro, e independencia entre el test y el estándar de oro. No habría sesgo de verificación, pues se aplicó el test y el estándar de oro a la totalidad de los pacientes). Por último, es importante señalar que el uso de curvas ROC trasciende el área de los tests diagnósticos, siendo frecuentemente empleadas para evaluar la capacidad predictora de modelos de regresión logística, ampliamente utilizados en investigación clínica y poblacional. Invitamos a los lectores a profundizar sus conocimientos referentes a esta valiosa herramienta estadística, de gran utilidad para la práctica clínica e investigación biomédica.

*Agradecimientos***.** Los autores agradecen a Luis Villarroel, Ph.D en Estadística (Pontificia Universidad Católica de Chile) por sus valiosas sugerencias.

Resumen

Una amplia gama de tests diagnósticos reportan sus resultados cuantitativamente, utilizando escalas contínuas. El análisis de curvas ROC *(receiver operating characteristic curve)* constituye un método estadístico para determinar la exactitud diagnóstica de estos tests, siendo utilizadas con tres propósitos específicos: determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta, evaluar la capacidad discriminativa del test diagnóstico, es decir, su capacidad de diferenciar sujetos sanos *versus* enfermos, y comparar la capacidad discriminativa de dos o más tests diagnósticos que expresan sus resultados como escalas continuas. En base a una investigación clínica real, el presente artículo ilustra el sustrato teórico subyacente a la construcción de curvas ROC, con el objetivo de ayudar a lectores e investigadores a interpretar correctamente sus resultados.

Referencias

1.- Cerda J, Cifuentes L. Uso de tests diagnósticos en la práctica clínica (Parte 1). Análisis de las propiedades de un test diagnóstico. Rev Chil Infectol 2010; 27: 205-8.
2.- Cifuentes L, Cerda J. Uso de tests diagnósticos en la práctica clínica (Parte 2). Aplicación clínica y utilidad de un test diagnóstico. Rev Chil Infectol 2010; 27: 316-9.
3.- Akobeng A. Understanding diagnostic tests 3: receiver characteristic characteristic curves. Acta Paediatr 2007; 96: 644-7.
4.- Altman D G, Bland J M. Diagnostic tests 3: receiver operating characteristic plots. Br Med J 1994; 309: 188.
5.- Mueller C, Huber P, Laifer G, Mueller B, Perrochoud A. Procalcitonin and the early diagnosis of infective endocarditis. Circulation 2004; 109: 1707-10.
6.- Durack D T, Lukes A S, Bright D K, Duke Endocarditis Service. New criteria for diagnosis of infective endocarditis: utilization of specific echocardiographic findings. Am J Med 1994; 96: 200-9.
7.- Hanley J A, McNeil B J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-43.
8.- DeLong E R, DeLong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837-45.
9.- Valenzuela L, Cifuentes L. Validez de estudios de tests diagnósticos. Rev Med Chile 2008; 136: 401-4.