Rev Esp Cardiol. 2011;64(9):788–794

Enfoque: Métodos contemporáneos en bioestadística (IV)

Medidas del rendimiento de modelos de predicción y marcadores pronósticos: evaluación de las predicciones y clasificaciones

Ewout W. Steyerberg[a,*], Ben Van Calster[a,b] y Michael J. Pencina[c]

[a] Department of Public Health, Erasmus MC, Rotterdam, Países Bajos
[b] Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Lovaina, Bélgica
[c] Department of Biostatistics, Boston University, y Harvard Clinical Research Institute, Boston, Massachusetts, Estados Unidos

Historia del artículo:
On-line el 16 de julio de 2011

Palabras clave:
Predicción
Clasificación
Modelo de regresión
Análisis de decisión

RESUMEN

Los modelos de predicción están adquiriendo cada vez mayor importancia en medicina y en cardiología. En la actualidad, hay un interés específico que se centra en las formas de mejorar los modelos con el empleo de nuevos marcadores pronósticos. Nuestro objetivo es describir las semejanzas y diferencias entre las distintas medidas del rendimiento de los modelos de predicción. Hemos analizado los datos de 3.264 individuos para predecir el riesgo de enfermedad coronaria a 10 años, según la edad, la presión arterial sistólica, la diabetes y el tabaquismo. Estudiamos específicamente el valor incremental de la adición a este modelo del colesterol unido a lipoproteínas de alta densidad.

Resaltamos que es preciso separar la evaluación de las predicciones —en las que las medidas de rendimiento tradicionales, como el área bajo la curva *receiver operating characteristic* y la calibración, resultan útiles— de la evaluación de las clasificaciones, para las que disponemos actualmente de otros parámetros estadísticos, como el *net reclassification index* y el beneficio neto.

Performance Measures for Prediction Models and Markers: Evaluation of Predictions and Classifications

Keywords:
Prediction
Classification
Regression model
Decision analysis

ABSTRACT

Prediction models are becoming more and more important in medicine and cardiology. Nowadays, specific interest focuses on ways in which models can be improved using new prognostic markers. We aim to describe the similarities and differences between performance measures for prediction models. We analyzed data from 3264 subjects to predict 10-year risk of coronary heart disease according to age, systolic blood pressure, diabetes, and smoking. We specifically study the incremental value of adding high-density lipoprotein cholesterol to this model.

We emphasize that we need to separate the evaluation of predictions, where traditional performance measures such as the area under the receiver operating characteristic curve and calibration are useful, from the evaluation of classifications, where various other statistics are now available, including the net reclassification index and net benefit.

Full English text available from: www.revespcardiol.org

INTRODUCCIÓN

Los modelos de predicción están adquiriendo una importancia creciente en la literatura médica. Actualmente disponemos de muchos modelos para la predicción de un diagnóstico (la presencia de una enfermedad) o un pronóstico (p. ej., la incidencia de enfermedad coronaria [EC]). La cuantificación del riesgo cardiovascular se realiza generalmente mediante ecuaciones de riesgo o gráficos de puntuación del riesgo que se han desarrollado a partir de estudios de cohorte amplios[1]. Las técnicas de modelización incluyen el modelo de riesgos proporcionales de Cox y el modelo paramétrico de Weibull[2].

Las funciones de riesgo de Framingham son uno de los ejemplos mejor conocidos de estos modelos de predicción[1,3] y han sido esenciales para individualizar las decisiones de tratamiento preventivo, por ejemplo sobre el uso del tratamiento con estatinas. Ahora el interés específico se centra en cómo se puede mejorar la predicción del riesgo con el empleo de los nuevos marcadores[4] identificados gracias a los avances tecnológicos en la investigación básica, incluidas la genómica, la proteómica y las técnicas de imagen no invasivas. Estos marcadores parecen prometedores para aproximarse a la medicina personalizada. Una cuestión importante es cómo evaluar la utilidad de un nuevo marcador para la toma de mejores decisiones, como dirigir el tratamiento con estatinas a los pacientes de mayor riesgo[5].

* Autor para correspondencia: Department of Public Health, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, Países Bajos.
Correo electrónico: e.steyerberg@erasmusmc.nl (E.W. Steyerberg).

0300-8932/$ – see front matter
doi:10.1016/j.recesp.2011.04.017

Abreviaturas

AUC: área bajo la curva *receiver operating characteristic*
B: beneficio producido por una clasificación positiva verdadera
BN: beneficio neto
D: daño producido por una clasificación falsa positiva
FP: número total de clasificaciones falsas positivas en el conjunto de datos
NRI: *net reclassification index*
ROC: *receiver operating characteristic*
VP: número total de clasificaciones positivas verdaderas en el conjunto de datos

Una condición básica que debe cumplir un nuevo marcador es la significación estadística, que generalmente se define mediante un valor de p bilateral $< 0,05$. Sin embargo, la significación estadística no implica trascendencia clínica o utilidad del marcador. De hecho, un biomarcador con una relación débil con el resultado de interés puede mostrar una asociación estadísticamente significativa si se examina un tamaño muestral suficientemente grande.

Nuestro objetivo en este artículo es describir las semejanzas y diferencias entre las distintas medidas del rendimiento de los modelos de predicción. Nos centramos específicamente en las medidas destinadas a cuantificar la mejora del rendimiento predictivo con la adición de un marcador a un modelo de predicción existente.

MÉTODOS Y RESULTADOS

Pacientes

El *Framingham Heart Study* se inició en 1948 con una cohorte de 5.209 individuos. En 1971, 5.124 participantes (hijos de los individuos de la cohorte inicial y de sus cónyuges) fueron incluidos en el *Framingham Offspring Study*. De ellos, 3.951 participantes de entre 30 y 74 años de edad acudieron al cuarto ciclo de exámenes de la cohorte del *Framingham Offspring* entre 1987 y 1992.

Según se ha descrito anteriormente, excluimos a los participantes con una EC conocida o de los que no se disponía de datos de los factores de riesgo estándar, con lo que quedaron 3.264 de los 3.951 para el presente análisis[5]. Los participantes estuvieron en seguimiento durante 10 años para identificar la aparición de EC (incluidos infarto de miocardio, angina de pecho, insuficiencia cardiaca y muerte por EC). En total, 183 individuos contrajeron una EC (5,6%). Estos datos constituyen un ejemplo que permite ilustrar los conceptos, más que llevar a cabo un análisis exhaustivo.

Análisis

Se elaboraron modelos de riesgos proporcionales de Cox con sexo, diabetes mellitus y tabaquismo como factores predictivos dicotómicos y edad, presión arterial sistólica y colesterol total como factores predictivos continuos. Las razones de riesgos fueron estadísticamente significativas para todos estos factores predictivos. La adición a este modelo del colesterol unido a lipoproteínas de alta densidad (cHDL) como factor predictivo continuo fue altamente significativa (razón de riesgos = 0,65; $p < 0,001$)[5].

Analizamos con mayor detalle la mejora del rendimiento del modelo como consecuencia de la inclusión del cHDL, mediante la comparación de dos conjuntos de predicciones de la probabilidad de riesgo de EC a 10 años: un conjunto de predicciones basadas en un modelo de riesgos proporcionales de Cox sin la inclusión del cHDL y un conjunto de predicciones basadas en un modelo con la inclusión del cHDL.

Medidas del rendimiento respecto a la calidad de las predicciones

Discriminación

Una medida clave de un modelo de predicción es su capacidad de diferenciar a los individuos que sufrirán el evento de interés de los que no; en nuestro caso, la aparición de EC frente a la ausencia de EC a los 10 años de seguimiento[6]. El área bajo la curva (AUC) *receiver operating characteristic* (ROC) es la medida más utilizada para cuantificar la capacidad de discriminación (tabla 1).

Tabla 1
Algunas medidas del rendimiento de los modelos de predicción: la evaluación de las predicciones se ha realizado con medidas distintas de las de la evaluación de la mejor clasificación con un marcador

Aspecto	Medida	Características
Evaluación de las predicciones		
Discriminación	AUC o estadístico *c*	AUC o *c* es un parámetro estadístico de orden de jerarquía; la interpretación consiste en la probabilidad de clasificación correcta para un par de pacientes con y sin el resultado evaluado
Calibración	Valor de intersección y pendiente de un modelo de recalibrado	Valor de intersección ($a/b = 1$), que refleja la calibración en general, o la diferencia entre la media de predicciones y la media de resultados. Pendiente de recalibración *(b)*, que refleja la media del efecto de los factores predictivos en el resultado
Evaluación de las clasificaciones		
Clasificación	Índice de Youden	Suma de sensibilidad y especificidad – 1
Utilidad clínica	BN y DCA	Fracción neta de los positivos verdaderos ganados mediante la toma de decisiones basada en las predicciones para un único umbral (BN) o para una gama de umbrales (DCA)
Evaluación del valor incremental con un marcador		
Aumento de la discriminación	Delta de AUC	El aumento de la discriminación suele ser una cifra pequeña
Reclasificación	NRI	Fracción neta de reclasificaciones en el sentido correcto mediante la toma de decisiones basadas en las predicciones realizadas con un marcador en comparación con las decisiones tomadas sin el marcador
Utilidad clínica	Diferencia en BN y DCA; NRI ponderado	Fracción neta de positivos verdaderos ganados con la toma de decisiones basada en predicciones realizadas con un marcador en comparación con las decisiones tomadas sin el marcador para un único umbral (BN) o en una gama de umbrales (DCA); ponderaciones según las consecuencias de las decisiones (BN y NRI ponderado)

AUC: área bajo la curva ROC; BN: beneficio neto; DCA: análisis de curva de decisión; NRI: n*et reclassification index*; ROC: *receiver operating characteristic*.

La curva ROC representa gráficamente la relación entre la sensibilidad (la tasa de positivos verdaderos, o sea, la probabilidad de EC en los clasificados como positivos) y 1 menos la especificidad (la tasa de falsos positivos, o sea, la probabilidad de ausencia de EC en los clasificados como negativos). Se calculan pares de valores de sensibilidad y especificidad para todos los posibles valores de corte para las probabilidades predichas del riesgo de EC a 10 años. Con un valor de corte bajo como el del riesgo del 0,1%, la sensibilidad es alta, pero la especificidad es mala. Un valor de corte del 5,6% corresponde a la incidencia de la EC (a veces se denomina «prevalencia»). A ese valor de corte, el modelo sin las lipoproteínas de alta densidad (HDL) tenía una sensibilidad del 74% y una especificidad del 65% (fig. 1). El modelo con las HDL daba mejores resultados a ese valor de corte (sensibilidad, 78%; especificidad, 66%). Un valor de corte más alto, como el del 20%, implicaba una sensibilidad inferior, pero con mayor especificidad (fig. 1).

El AUC es igual a la probabilidad de que, entre dos individuos dados (uno que sufre una EC en el seguimiento de 10 años y otro que no), el modelo asigne una probabilidad de EC más alta al primero de ellos. El AUC para el modelo sin las HDL en comparación con el modelo con HDL fue de 0,762 (intervalo de confianza [IC] del 95%, 0,73-0,794) frente a 0,774 (0,742-0,806). Esta diferencia de 0,012 es difícil de interpretar, pero la mayoría de los investigadores podrían considerarla pequeña.

Calibración

Otra dimensión importante en la calidad de las predicciones es la calibración, es decir, la coincidencia entre las probabilidades predichas y las frecuencias observadas del evento de interés[6]. Por ejemplo, en los individuos para los que se predice un riesgo del 5% del evento de interés, en promedio, 5/100 deberían presentar el evento en cuestión. Una forma de estudiar el calibrado es representar gráficamente una función de los eventos observados frente a las probabilidades predichas, por ejemplo con el empleo de una curva *loess* (fig. 2)[6]. En el caso ideal[2], se obtiene una línea a 45°, con una pendiente de 1 y un punto de intersección de 0. La pendiente y el punto de intersección pueden calcularse en un modelo de regresión que tiene en cuenta una transformación de las probabilidades predichas como único factor predictivo del resultado. En nuestro caso, observamos una calibración casi perfecta del modelo logístico para la EC a 10 años, con el *logit* de las probabilidades previstas obtenidas a partir del modelo de Cox (fig. 2).

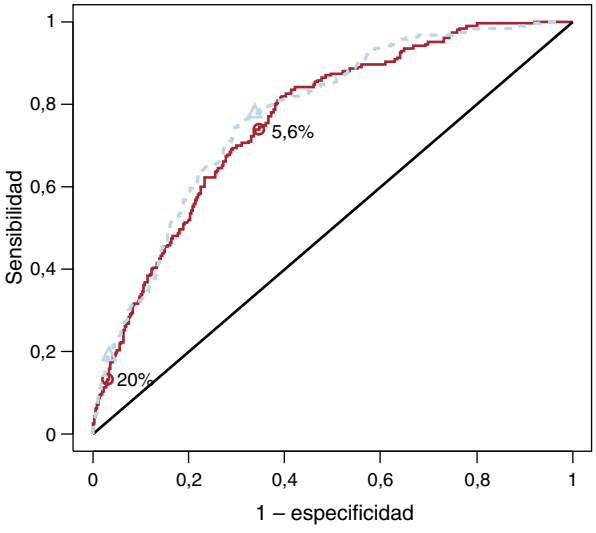

Figura 1. Curvas *receiver operating characteristic* para los modelos de predicción del riesgo de enfermedad coronaria a 10 años, basadas en 3.264 individuos. Las áreas fueron de 0,762 frente a 0,774 para el modelo sin las lipoproteínas de alta densidad y el modelo con las lipoproteínas de alta densidad, respectivamente. Se muestran dos valores de corte: 5,6% es la media de incidencia de enfermedad coronaria a 10 años, y 20% es un valor de corte clínicamente relevante para definir a individuos con riesgo alto.

Evaluación gráfica de la calidad de las predicciones

En la figura 2 mostramos también las distribuciones de las probabilidades predichas en los individuos con y sin EC para visualizar la discriminación[7]. Hay un considerable solapamiento entre estas distribuciones, lo cual ilustra el significado de los valores de AUC de 0,76 y 0,77. Las medidas que resume este gráfico pueden abreviarse como *a, b* y *c: a* indica el valor de intersección, la calibración en general; *b,* la pendiente de recalibración y *c,* el AUC[2].

Determinación del valor de corte para la clasificación

La curva ROC tiene en cuenta todos los valores de corte consecutivos para definir un grupo de riesgo elevado frente a un grupo de riesgo bajo. Hay diversas maneras de determinar un valor de corte óptimo. Comentaremos un enfoque basado en los datos y un enfoque de análisis de decisión (o «basado en utilidad»).

Valor de corte basado en los datos

Una medida bien conocida para clasificar el rendimiento es el índice de Youden, que se define como la sensibilidad + especificidad − 1[8]. El índice de Youden es máximo en el ángulo superior izquierdo de la curva ROC. Por consiguiente, podríamos buscar el valor de corte correspondiente a ese punto. Es interesante señalar que el punto situado en el ángulo superior izquierdo corresponde al uso de la incidencia del resultado como valor de corte para la probabilidad predicha, si el modelo de predicción está bien calibrado y la curva ROC es cóncava[9]. En nuestro caso, este valor de corte es de 183/3.264 = 5,6% (fig. 1).

Valor de corte de análisis de decisión

El análisis de decisión toma el contexto clínico como punto de partida. Se considera formalmente la utilidad, o satisfacción relativa, de la consecuencia de una clasificación verdadera o falsa[10]. En el caso de la prevención de la EC, un valor de corte ampliamente aceptado para definir un grupo de alto riesgo es el del 20%. Formalmente, este valor de corte del 20% implica que la utilidad de las clasificaciones falsas positivas es 4 veces inferior a la de las clasificaciones positivas verdaderas, es decir, (100 – 20)/20[7]. Una clasificación falsa positiva implica un sobretratamiento: un individuo que no sufrirá una EC en 10 años es tratado, por ejemplo, con estatinas. El daño se pondera como 4 veces menos importante que el beneficio de una clasificación positiva verdadera (un individuo que sufrirá una EC en 10 años es tratado con estatinas). Expresado en una fórmula, la probabilidad del valor de corte es igual al cociente entre daño (D) y beneficio (B):

$$\text{Probabilidad (valor de corte)} = D/B.$$

Un valor de corte del 50% (probabilidad = 1) implica un cociente D:B de 1:1; un valor de corte del 20% (probabilidades = 1/4) implica un cociente de 1:4. Un valor de corte del 5,6% maximiza la suma de sensibilidad y especificidad, pero implica que consideramos los

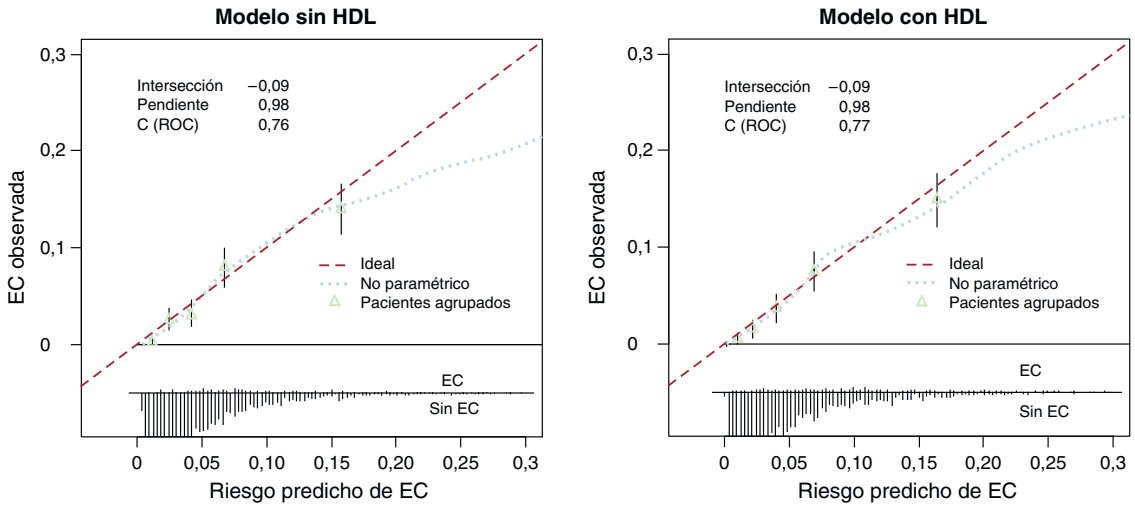

Figura 2. Gráficos de validación para el modelo sin las lipoproteínas de alta densidad y con las lipoproteínas de alta densidad para la predicción de la enfermedad coronaria en 10 años de seguimiento. El «valor de intersección» hace referencia a la calibración general, y la «pendiente» se refiere a la pendiente de calibración para las predicciones. «C (ROC)» se refiere al área bajo la curva *receiver operating characteristic*. La línea ideal de 45° tiene un valor de intersección de 0 y una pendiente de 1. Los triángulos indican los resultados para los quintiles de las predicciones, con intervalos de confianza del 95%. Las puntas en la parte inferior indican las predicciones para los individuos con y sin enfermedad coronaria. EC: enfermedad coronaria; HDL: lipoproteínas de alta densidad; ROC: *receiver operating characteristic*.

falsos positivos casi 20 veces menos importantes que los positivos verdaderos (0,056/0,944).

Medidas del rendimiento respecto a la calidad de las clasificaciones

Curvas receiver operating characteristic con 1 valor de corte

En vez de considerar todos los posibles valores de corte en las curvas ROC, podemos construir también curvas ROC utilizando un solo valor de corte basado en los datos (fig. 3A) o basado en un análisis de decisión (fig. 3B). Las AUC son de 0,696 y 0,719 para el valor de corte del 5,6% y de 0,55 y 0,579 para el valor de corte del 20% en los modelos sin las HDL y con las HDL, respectivamente. Es interesante señalar que el aumento del AUC con la adición al modelo de predicción de las HDL se ha incrementado ahora (pasando de 0,012 para todos los valores de corte a 0,023 y 0,029 para los valores de corte del 5,6 y el 20%, respectivamente).

Reclasificación

Cook ha reconocido que el valor incremental de un marcador se expresa como los cambios que se producen en la clasificación del riesgo cuando se consideran las probabilidades predichas del marcador en el modelo predictivo[11]. Por ejemplo, tener en cuenta las HDL conduce a una reclasificación del 9,8% de los individuos con el empleo del valor de corte del 5,6%. Esta cifra próxima al 10% tiene más impacto que el aumento de 0,01 del AUC para todos los valores de corte o que el aumento de 0,02 con el uso del valor de corte del 5,6%.

Reclasificación neta

Pencina et al[5] han señalado que no deberíamos tener tan en cuenta la reclasificación en todos los pacientes, sino centrarnos en la reclasificación en el sentido correcto, es decir, un clasificación en un riesgo superior en los individuos con EC y en un riesgo

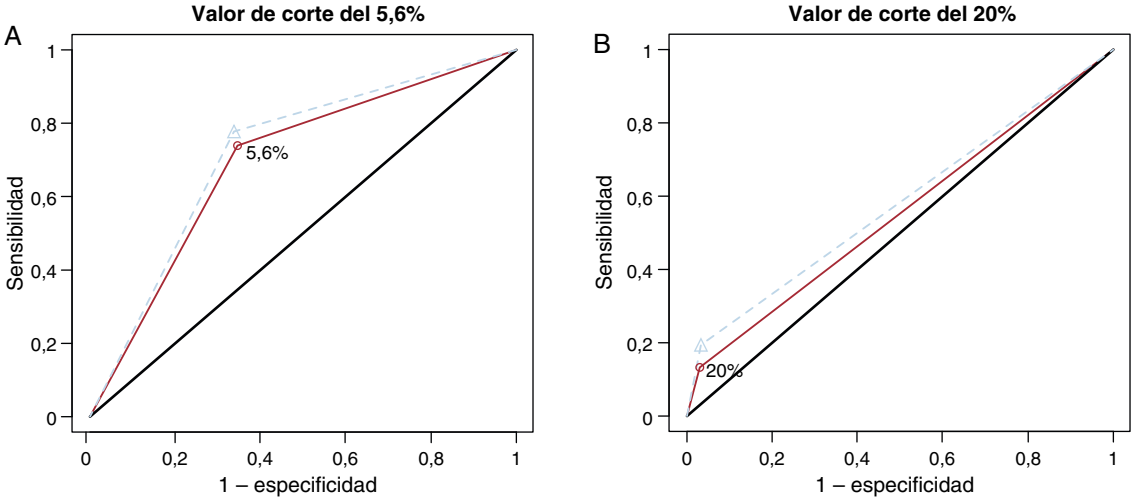

Figura 3. Curvas *receiver operating characteristic* con valores de corte únicos del 5,6 (A) y el 20% (B). El área bajo las curvas *receiver operating characteristic* es de 0,696 y 0,719 para el valor de corte del 5,6% y de 0,55 y 0,579 para el valor de corte del 20%, en el modelo con las lipoproteínas de alta densidad y sin las lipoproteínas de alta densidad, respectivamente.

Tabla 2
Reclasificación en 3.264 individuos con o sin un evento de enfermedad coronaria en un plazo de 10 años de seguimiento

	Modelo sin las HDL	Modelo con las HDL	
		≤ 5,6%	> 5,6%
Ausencia de EC (n = 3.081)	≤ 5,6%	1.872	142[a]
	> 5,6%	166[b]	901
EC (n = 183)	≤ 5,6%	38	10[b]
	> 5,6%	3[a]	132

EC: enfermedad coronaria; HDL: lipoproteínas de alta densidad.
[a] Reclasificaciones en sentido erróneo.
[b] Reclasificaciones en sentido correcto.

inferior en los individuos sin EC. Con el empleo del valor de corte del 5,6%, esta reclasificación neta es de 7/183 (3,8%) para los individuos con EC, y de 24/3.081 (0,8%) para los individuos sin EC (tabla 2). La suma de estas cifras corresponde al índice de reclasificación neta (*net reclassification index* [NRI]): 4,6% (IC del 95%, 0,6-8,6%). Para el valor de corte del 20%, NRI = 5,8% (1,4-10,3%).

Beneficio neto

Ya en 1884, Peirce[12] afirmó que la calidad de las clasificaciones puede expresarse mediante la suma ponderada de las clasificaciones positivas verdaderas: el beneficio neto (BN). El BN compensa las clasificaciones falsas positivas dándoles una ponderación *w:*

$$BN = (VP - w\,FP)/N$$

donde VP es el número de clasificaciones positivas verdaderas, FP es el número de clasificaciones falsas positivas y N, el número total de individuos.

Si $w = 1$, FP y VP se ponderan por igual. Como se ha comentado antes, esto implica una probabilidad de 1:1 para el cociente D:B. De hecho, w es el cociente D:B. Así pues, un cociente D:B de 1:4 implica un valor de corte del 2% y una ponderación de 0,25 para las clasificaciones FP respecto a las clasificaciones VP, y un valor de corte del 5,6% implica $w = 0{,}056/0{,}944 = 0{,}059$.

Teniendo en cuenta las cifras indicadas en la tabla 2, el BN en el modelo sin las HDL se calcula de la siguiente forma: VP = 3 + 132 = 135; FP = 166 + 901 = 1.067; $w = 0{,}056/0{,}944 = 0{,}059$, y N = 3.264. Esto lleva a un BN de $(135 - 0{,}059 \times 1.067)/3.264 = 2{,}21\%$. Para el modelo con las HDL, el BN es superior: $(142 - 0{,}059 \times 1.043)/3.264 = 2{,}47\%$. El aumento de las clasificaciones VP es de 10 – 3 = 7, y la disminución de las clasificaciones FP es de 166 – 142 = 24. Esto explica el aumento del BN de $(7 + 0{,}059 \times 24)/3.264 = 0{,}26\%$. Esta cifra puede interpretarse como un aumento neto de las clasificaciones positivas verdaderas, es decir, se identifican 2,6 eventos de EC verdaderos más por cada 1.000 individuos con el mismo número de clasificaciones FP[13]. Esto equivale a decir que es preciso determinar las HDL en 1/0,26% = 385 individuos para identificar un VP más, utilizando un valor de corte del 5,6%.

Curvas de decisión

El valor de corte para la aplicación clínica de un modelo de predicción a menudo no se define de manera precisa. La ponderación relativa de daños y beneficios puede no ser conocida a causa de la falta de datos científicos o debido a apreciaciones diferentes de distintos médicos y pacientes. Por este motivo, Vickers y Elkin[13] propusieron utilizar una gama de valores de corte y calcular el BN para estos distintos valores. El resultado puede representarse gráficamente en una curva de decisión (fig. 4).

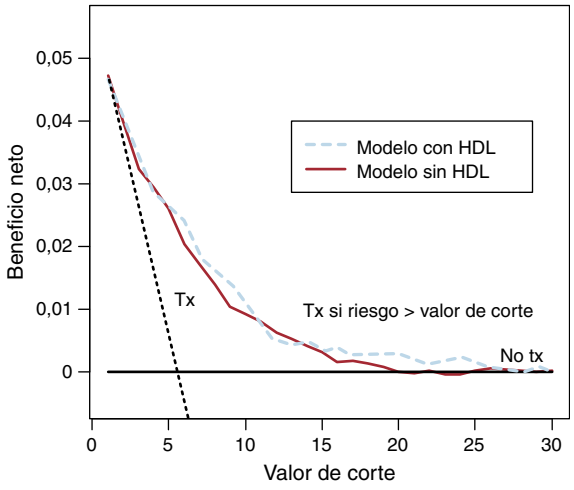

Figura 4. Curva de decisión para el modelo con las lipoproteínas de alta densidad y sin las lipoproteínas de alta densidad para la predicción de la enfermedad coronaria en un seguimiento de 10 años. La línea punteada pequeña indica el beneficio neto para tratarlos a todos, mientras que la línea horizontal corresponde a no tratar a ninguno. Estas dos líneas sirven de referencia para las líneas del beneficio neto en los modelos con o sin lipoproteínas de alta densidad. HDL: lipoproteínas de alta densidad; Tx: tratamiento.

Observamos que se obtiene un BN pequeño al añadir al modelo las HDL para valores de corte entre el 5 y el 25%.

Más valores de corte para la clasificación

En la enfermedad cardiovascular es frecuente el uso de tres grupos de riesgo[1,5]. Un grupo de bajo riesgo puede definirse por un riesgo < 6%, un grupo de alto riesgo que requiere un tratamiento preventivo intensivo se define por un riesgo > 20% y los demás individuos se clasifican como de riesgo intermedio y necesitan recomendaciones de estilo de vida, por ejemplo. Podemos calcular diversas medidas para estos dos valores de corte, como el AUC y el NRI. No es posible calcular directamente el BN, dado que este se define para 1 valor de corte.

También podemos considerar toda la gama de valores de corte para la reclasificación en un NRI de menos categorías. El NRI (> 0) se define como un cambio en el sentido correcto para cualquier valor de corte considerado[14]. Este cálculo debe considerarse de nuevo por separado para los individuos con y sin EC. En nuestro caso, el 62% de los 183 individuos con EC tuvieron predicciones superiores en el modelo con las HDL y el 38% tuvo predicciones inferiores, con lo que el NRI para los eventos era del 24,6%. Para los 3.081 individuos sin EC, el 53% tuvo predicciones inferiores con el modelo con las HDL y el 47%, predicciones superiores, con un NRI del 5,6%. El NRI (> 0) fue de 0,3. Estos patrones pueden evaluarse también gráficamente, comparando las predicciones con o sin la inclusión de las HDL en el modelo, en un gráfico de reclasificación (fig. 5)[7,14,15]. Señalamos aquí que hay un número ligeramente superior de puntos por debajo de la línea de 45° para los individuos sin EC y que hay un número sustancialmente superior de puntos situados por encima de la línea de 45° para los individuos con EC.

Interrelaciones

Si utilizamos un único valor de corte, la AUC = (sensibilidad + especificidad)/2. El aumento de AUC (o Δ_{AUC}) es pues de $0{,}5 \times (\Delta_{sensibilidad} + \Delta_{especificidad})$. El NRI[14] en este caso de dos categorías es $\Delta_{sensibilidad} + \Delta_{especificidad}$, o $2 \times \Delta_{AUC}$. Dado que el índice de Youden = (sensibilidad + especificidad) – 1, Δ_{Youden} es $\Delta_{sensibilidad} + \Delta_{especificidad}$, igual a NRI. De hecho, el aumento de AUC

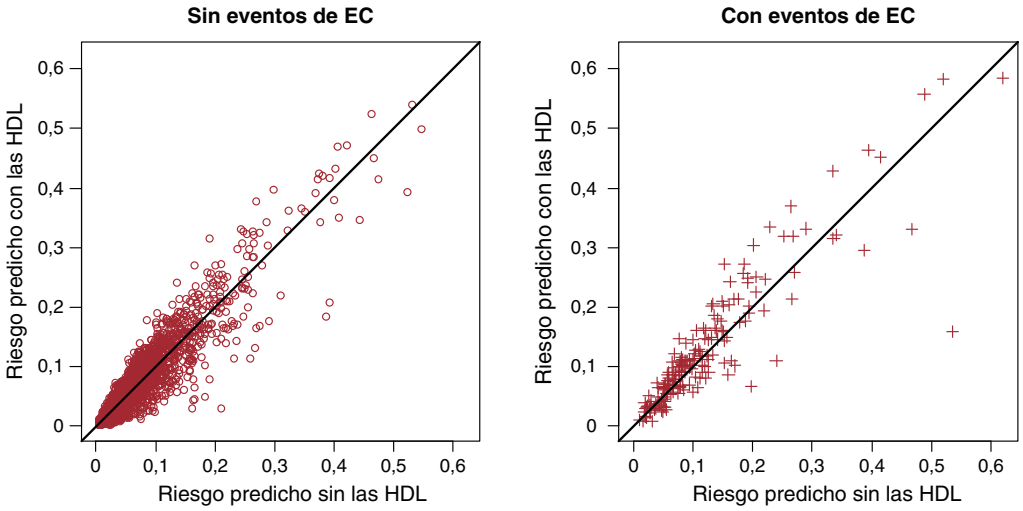

Figura 5. Gráfico de reclasificación. EC: enfermedad coronaria; HDL: lipoproteínas de alta densidad.

fue de 0,023 para el valor de corte del 5,6%, mientras que el NRI y el índice de Youden fue de 0,046. Así pues, parece claro que el NRI es una cifra superior al aumento de AUC.

El NRI (> 0) está relacionado con el Δ_{AUC} para todos los valores de corte posibles. Las comparaciones utilizadas en el cálculo de NRI (> 0) están entre los dos modelos de predicción (con y sin el marcador), pero dentro de los grupos del evento (con EC, sin EC). El Δ_{AUC} se basa en comparaciones apareadas entre los grupos del evento (con EC, sin EC) dentro de los dos modelos de predicción[14].

El BN es una suma ponderada de la sensibilidad (fracción de VP) y 1 – especificidad (fracción de FP). Si el valor de corte es la incidencia del resultado, el NRI con dos categorías es igual a Δ_{BN}/incidencia. La incidencia a 10 años de EC fue del 5,6%. De hecho, el aumento del BN fue del 0,26% para el valor de corte del 5,6%, mientras que el NRI fue del 4,6% (= 0,0026/0,056). Así pues, parece claro que el NRI es una cifra muy superior al aumento del BN. Recientemente se ha propuesto una variante ponderada del NRI, que se comporta de manera similar al BN como medida de resumen de la utilidad de añadir un marcador a un modelo[14].

DISCUSIÓN

Hemos mostrado de qué forma pueden utilizarse diversas medidas interrelacionadas para indicar el rendimiento de un modelo de predicción. Hemos ilustrado estas medidas con un modelo de riesgo elaborado para predecir la incidencia de EC a 10 años, con o sin el uso del cHDL como marcador de riesgo. Hemos diferenciado la evaluación de las predicciones —para lo que son útiles las medidas del rendimiento tradicionales como el AUC y la calibración— de la evaluación de las clasificaciones y la contribución de los nuevos marcadores, para lo que disponemos actualmente de otros varios parámetros estadísticos, como el NRI y el BN[5,7,13,14].

La distinción entre un modelo de predicción y una regla de predicción no está clara en la mayor parte de la actual literatura diagnóstica y pronóstica. El elemento clave es que, para pasar de un modelo de predicción a una regla de predicción, es necesaria la definición de un umbral de decisión o valor de corte[16]. «Modelo de predicción» y «regla de predicción» no son, pues, sinónimos. En una regla de predicción, los pacientes con predicciones situadas por encima y por debajo del umbral se clasifican como positivos y negativos, respectivamente. Es de destacar que el AUC y el NRI (> 0) evalúan modelos y no reglas. Sin embargo, un buen modelo es el primer paso para elaborar una buena regla.

El umbral para una regla debe ser apropiado, teniendo en cuenta las consecuencias (o *utilidades*) de la decisión[10]. A menudo una clasificación falsa positiva (sobrediagnóstico) es menos ponderada en el contexto médico que una clasificación falsa negativa (infradiagnóstico de la enfermedad)[16]. En el caso en estudio, el umbral de decisión del 20% refleja una ponderación relativa de 1:4 para las clasificaciones falsas positivas y positivas verdaderas. Una vez utilizada una ponderación relativa para definir el umbral de decisión, es lógico ser coherente y aplicar también esta ponderación relativa en la evaluación de la calidad de las decisiones. Este principio se sigue en la definición del BN y en el NRI ponderado[14], así como en las medidas relativas como la utilidad relativa[17]. El NRI de dos categorías no concuerda generalmente con el Δ_{BN} o la utilidad relativa. Tan sólo si el umbral de decisión es igual a la incidencia del resultado se obtienen resultados coincidentes con el NRI y el Δ_{BN}.

El NRI se ha popularizado rápidamente como medida de resumen del valor predictivo de un marcador. Obsérvese que las publicaciones metodológicas han resaltado siempre la consideración de los componentes individuales del NRI[5,14], es decir, el NRI para los eventos y el NRI para la ausencia de eventos, como se muestra en la tabla 2.

Una de las razones de la popularidad del NRI puede estar en que el número absoluto se presenta a menudo en forma de porcentaje y, por lo tanto, es sustancialmente superior al aumento del AUC. En nuestro ejemplo, el Δ_{AUC} para todos los valores de corte fue de 0,012 (fig. 1), mientras que el NRI fue de +4,6% para un valor de corte del 5,6%. Así pues, el NRI es casi 4 veces el Δ_{AUC}. Sin embargo, para realizar una comparación justa, habría que considerar el valor de corte del 5,6% también para el Δ_{AUC}, que fue del 2,3%. Con ello aparece la relación matemática simple de que el NRI = 2 veces el Δ_{AUC}[14]. Pueden obtenerse valores aún mayores del NRI si se consideran todos los valores de corte (NRI [> 0] + 30%).

Otra razón de la popularidad del NRI es que se considera que el AUC «no es sensible» a los aumentos del valor predictivo de un marcador[11]. En una evaluación reciente, se observó una potencia estadística limitada para el Δ_{AUC} en comparación con un cociente de probabilidades o con la prueba de Wald para la adición de un marcador a un modelo de regresión[18]. Sin embargo, estos autores llegaron a la conclusión de que las comparaciones de los valores de AUC continuaban siendo útiles para la evaluación inicial de si un

nuevo predictor puede tener relevancia clínica. No hay motivo alguno para presumir que la potencia estadística del NRI sea mejor que la de la prueba de cociente de probabilidades; por el contrario, la clasificación en categorías comporta una pérdida de información predictiva y debería conducir a una potencia estadística inferior a la de una prueba que incluya toda la gama de probabilidades predichas. En nuestra opinión, la principal cuestión en la evaluación del rendimiento no es la potencia estadística, sino la interpretación de la calidad de un modelo y la mejora de este con los marcadores.

Limitaciones

Nuestro estudio tiene varias limitaciones. No utilizamos métodos específicos para los datos de supervivencia, a pesar de que no se dispuso de un seguimiento completo hasta los 10 años de todos los individuos. Se supuso simplemente que los pacientes censurados no tenían EC. Existen métodos para calcular el AUC (en forma de concordancia, o estadístico *c*) y el NRI para datos de supervivencia[14,19]. Además, no evaluamos el rendimiento como estudio de validación en datos independientes. Es frecuente que los estudios iniciales de modelos de predicción y marcadores muestren resultados prometedores, y las evaluaciones posteriores sean desalentadoras. La validación interna con validación cruzada o remuestreo *(bootstrapping)* constituye una exigencia mínima[20]. El tamaño muestral relativamente grande (n = 3.264 individuos; 183 eventos) probablemente hizo que el optimismo estadístico fuera bajo en nuestro caso (sin riesgo de sobreajuste), pero sería necesaria una validación externa.

Tras la validación y la determinación del valor predictivo, es necesario plantear estudios prospectivos de impacto para evaluar el valor de los modelos de predicción y los marcadores para la mejora de la evolución de los pacientes[16]. En primer lugar, podemos estudiar si un modelo con un marcador influye en la toma de decisiones médicas en comparación con un modelo sin el marcador. Si la decisión que se toma respecto a nuevos estudios diagnósticos o tratamientos no es diferente, no pueden mejorarse los resultados obtenidos en el paciente. El estudio ideal sería un ensayo aleatorizado sobre el impacto del aporte del valor del marcador en la evolución del paciente (morbilidad, mortalidad, calidad de vida), tomando los parámetros del proceso (pruebas diagnósticas, tratamientos administrados) como variables de valoración intermedias en el estudio[4]. Dado que con frecuencia los ensayos aleatorizados pueden no ser factibles a causa de la financiación necesaria para la investigación y el tamaño muestral requerido, puede ser pertinente también una modelización de análisis de decisión formal[21]. En estos modelos podemos combinar estimaciones del rendimiento del modelo de predicción con y sin el marcador con la evidencia disponible sobre la efectividad del tratamiento. El tratamiento podría aplicarse entonces de manera más apropiada a quienes lo necesitan.

CONCLUSIONES

En resumen, nosotros recomendamos la regla de *«a, b, c»* para la evaluación de las predicciones, en la que *a* (el punto de intersección) y *b* (la pendiente) se refieren a la calibración y *c*, al AUC (fig. 2). Para la evaluación de las clasificaciones y el valor de un marcador, el Δ_{AUC}, los componentes de eventos y ausencia de eventos del NRI, el NRI (> 0), el NRI ponderado y el BN son medidas de resumen apropiadas.

FINANCIACIÓN

Ewout Steyerberg contó con el apoyo de la *Netherlands Organization for Scientific Research* (subvención 9120.8004) y del *Center for Translational Molecular Medicine* (proyecto PCMM). Ben Van Calster recibe una subvención de formación posdoctoral de la Fundación de Investigación – Flanders (FWO).

CONFLICTOS DE INTERESES

Ninguno

BIBLIOGRAFÍA

1. Pencina MJ, D'Agostino RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119:3078–84.
2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837–47.
4. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119:2408–16.
5. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157–72.
6. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
8. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–5.
9. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991;11:95–101.
10. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109–17.
11. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928–35.
12. Peirce CS. The numerical measure of success of predictions. Science. 1884;4:453–4.
13. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
14. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11–21.
15. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008;168:2304–10.
16. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201–9.
17. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538–42.
18. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Method. 2011;11:13.
19. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Ann Intern Med. 2010;152:195–7.
20. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
21. Henriksson M, Palmer S, Chen R, Damant J, Fitzpatrick NK, Abrams K, et al. Assessing the cost effectiveness of using prognostic biomarkers with decision models: case study in prioritising patients waiting for coronary artery surgery. BMJ. 2010;340:b5606.