Scores de gravedad y complejidad en cirugía cardíaca. Usos y limitaciones

José M. Cortina Romero

Servicio de Cirugía Cardíaca. Hospital 12 de Octubre. Madrid. España.

La utilización de modelos predictivos para la estimación del riesgo quirúrgico en cirugía cardíaca, y especialmente en cirugía de revascularización coronaria, se ha convertido en los últimos años en una práctica habitual en el quehacer diario de cirujanos cardíacos y cardiólogos.

Hasta tal punto ha sido así que su uso para obtener estimaciones objetivas de mortalidad operatoria en cirugía coronaria se considera, en la revisión de las guías de práctica clínica de la AHA/ACC[1] de 2004, una recomendación clase IIa, con un grado de evidencia C.

De forma explícita, esta recomendación se refiere al uso de sistemas de predicción para realizar estimaciones preoperatorias del riesgo, que sirvan a médicos y enfermos a ponderar el equilibrio riesgo-beneficio del procedimiento quirúrgico que se plantea. Este uso, para estimaciones individuales preoperatorias, es sólo uno de los posibles de estos sistemas. Esta aplicación quizá sea la más intuitiva y, por ello, la que más demandan los clínicos. Sin embargo, es obligatorio recordar que, en su origen, el desarrollo de estos sistemas estuvo dirigido a conseguir estimaciones globales de resultados sobre series de pacientes y no sobre casos determinados.

Para explicar esto es necesario recurrir al origen del desarrollo de los diversos modelos. La existencia de sistemas de predicción y ajuste del riesgo en cirugía cardíaca es antigua y se remonta al CASS[2]. Pero lo que disparó su uso, tal como se concibe hoy, fue la publicación por parte de la Health Care Financing Administration (HCFA), en marzo de 1986, de los resultados «crudos», es decir, sin ajuste, de mortalidad en los hospitales que operaban enfermos de MEDICARE.

Esto originó una toma de posición[3] por parte de la Society of Thoracic Surgeons (STS) de Estados Unidos, que consideró que el uso de datos de mortalidad sin el adecuado ajuste a los factores de riesgo, era inapropiado y equívoco. A partir de ese momento comenzaron a surgir sistemas orientados a ponderar los resultados en función de la severidad de la enfermedad y de la presencia de morbilidad asociada.

METODOLOGÍA

No es el propósito de este comentario la descripción de cómo se construyen, evalúan y validan los modelos predictivos de los que se derivan los *scores*[4]. No obstante, conviene recordar que se debe demandar, al modelo por el que se opte, la máxima «robustez» metodológica en su construcción[5].

Sucintamente, el desarrollo de los modelos comienza por la definición precisa de la variable de salida o resultado, habitualmente la muerte hospitalaria, y del análisis de los posibles factores implicados en ella. Aunque pueda sorprender, la definición precisa de las variables es uno de los puntos más conflictivos. Incluso la muerte puede definirse de formas distintas. La definición imprecisa de algunas variables fue uno de los puntos débiles de uno de los modelos pioneros[6].

Otro motivo de controversia sistemática es el número de variables que debe manejar un modelo determinado. La visión de la práctica clínica habitual lleva a la idea intuitiva de que cuantas más variables, mejor se ajustará a la realidad. Esto, que esencialmente es verdad, en realidad sólo es aplicable cuando usamos los modelos para estimaciones individuales. Se ha contrastado que con un número relativamente pequeño de variables clínicas esenciales o «nucleares» se pueden construir modelos estadísticamente «robustos» y útiles para predicciones sobre series globales. Se ha documentado que la adición de nuevas variables a partir de un número determinado sólo incrementa de manera

VÉASE ARTÍCULO EN PÁGS. 515-22

Correspondencia: Dr. J.M. Cortina Romero.
Servicio de Cirugía Cardíaca. Hospital 12 de Octubre.
Avda. Córdoba, s/n. 28041 Madrid. España.
Correo electrónico: jcortina.hdoc@salud.madrid.org

Full English text available at: www.revespcardiol.org

marginal la capacidad predictiva. Hay que recordar que una condición imprescindible para el uso prospectivo de estos modelos es la aplicación de los *scores* a todos y cada uno de los casos susceptibles, sin excepción. Evidentemente esto resulta más sencillo cuando no es necesario computar muchas variables y las que se emplean tienen definiciones precisas.

Un punto de discusión interesante que conviene recordar, por su trascendencia, es la diferente capacidad predictiva de los modelos derivados de datos administrativos frente a los derivados de bases de datos clínicas. En general, y como es fácil intuir, estos últimos suelen mostrar mejor capacidad predictiva.

Una vez construido un modelo, se requiere una fase de validación que implica una serie de aspectos distintos que han de determinar si el modelo es fiable y «robusto». La interpretación habitual de lo que implica validación en el uso diario se refiere a validación predictiva y es sólo uno de los puntos que hay que considerar, aunque hay otros no menos importantes. La validación predictiva tiene dos aspectos que se han divulgado extensamente: calibración y discriminación. La calibración evalúa el modelo en su capacidad para la predicción de la mortalidad global y en los diferentes estratos de riesgo. La discriminación[7] es una medida de cómo el modelo es capaz de predecir bien un resultado determinado. Desde este punto de vista y a título orientativo, la cualificación de la capacidad de discriminación de los modelos depende del valor del área bajo la curva ROC.

Así, excelente discriminación serían valores mayores de 0,97. Muy buena estaría en el rango de entre 0,93 y 0,96; buena discriminación, entre 0,75 y 0,92, y por debajo de 0,75 serían modelos deficientes en su capacidad de discriminación.

USOS Y LIMITACIONES

A la hora de responder a interrogantes como qué modelo usar o qué limitaciones tienen, y otros muchos, considero que debemos tener en mente tres tipos de aplicaciones de estos *scores,* que están relacionadas pero que son distintas. Éstas serían: el uso para estimaciones individuales ante un enfermo determinado, el uso como descriptores del *case-mix* de las poblaciones que atendemos y, por último, como herramientas de control y gestión de la calidad prestada.

Estimaciones individuales

Es obligatorio recordar que los *scores* no se desarrollaron con este objetivo. Aunque tengan una capacidad de discriminación elevada, nunca será de 1. Por este motivo, su uso ante un paciente determinado sólo tiene una utilidad orientativa. Podemos estimar un riesgo determinado, pero jamás podremos predecir el resultado final en ese caso. Dicho de otra manera, modelos con una muy buena capacidad predictiva pueden estimar con poco error una mortalidad de 5 entre 100 pacientes, pero no serán capaces de determinar qué 5 pacientes morirán. Como recomiendan las guías de la AHA/ACC son útiles en el proceso de toma de decisión terapéutica. Realmente siguen siendo sorprendentes las divergencias en un sentido u otro entre estimaciones subjetivas del riesgo y las que nos proporcionan los *scores* ante un caso determinado.

Para este uso concreto, la recomendación más lógica sería el uso de modelos derivados de la experiencia del centro donde se va a realizar la técnica, pero, aunque hay grupos que han desarrollado su modelo predictivo, no existe tal grado de proliferación.

Para este tipo de uso deberíamos utilizar modelos logísticos que contemplen el mayor número de variables posibles y que computen todo el perfil clínico de un paciente determinado. Tanto el modelo actual de Bernstein y Parsonnet[8] como el que está disponible en la web de la STS (www.sts.org) se aproximan a estos requerimientos.

El consejo médico sobre un determinado procedimiento de un riesgo muy elevado es una decisión médica difícil. Es cierto que los pacientes con mayor riesgo son los que más se suelen beneficiar de los procedimientos cuando sobreviven. Pero no es menos cierto que hay cifras de riesgo que suponen, en la práctica, una mínima o nula posibilidad de supervivencia. En estos casos el consejo terapéutico es bastante complejo.

Descriptores del *case-mix* de las poblaciones

Una virtud de los sistemas de puntuación de este tipo es resumir en un número todo el perfil clínico de un paciente determinado, incluyendo los datos sobre la severidad de la enfermedad principal y de las patologías asociadas.

Esta facilidad es la que permite realizar una evaluación sencilla de las características globales de una población determinada, es decir, de su *case-mix.* De esta forma se puede comparar poblaciones de grupos, hospitales o incluso países distintos. Por la misma razón, se puede comparar la evolución en el tiempo de la casuística dentro de una misma institución. Como ejemplo de este uso es interesante la evolución del *case-mix* en el trabajo de García Fuster et al[9], en el que se evidencia un empeoramiento significativo de la gravedad de la población entre el primer y el segundo trienios, que posteriormente se estabiliza aunque con un discreto empeoramiento no significativo.

Para este tipo de uso como descriptor de poblaciones es obvio reseñar que no es obligatorio que el *score* se derive de una experiencia actual. Es evidente que, en el momento en que se usan como herramientas de medida, deberían permanecer estables en el tiempo, puesto que si no las comparaciones serían imposibles.

Como limitaciones para este uso de los *scores* creo que hay que resaltar dos especialmente importantes. En primer lugar, una que ya se demostró en el estado de Nueva York tras la publicación de los resultados de mortalidad por centro y por cirujano. Es la posible tendencia a sobrecargar artificiosamente el *case-mix* de los enfermos, sobre todo en presencia de variables de definición imprecisa. Lógicamente, el peso resultante no sería el auténtico que le correspondería con una aplicación estricta. Las únicas vías para neutralizar esto son el uso exclusivo de variables precisas e indiscutibles, la aplicación de los *scores* por agentes externos y la auditación sistemática del proceso de recogida de información.

La otra limitación que tienen como descriptores es su incapacidad para detectar variaciones en los criterios de indicación quirúrgica entre grupos o dentro de un mismo grupo. Así, variaciones del *case-mix* pueden traducir diferencias en la selección de enfermos sin que necesariamente tenga que haber diferencias en las características de las poblaciones con necesidad de ser atendidas.

Control de calidad de la actividad

Se puede decir que éste es el uso y la aplicación fundamental de los *scores* de riesgo. Su propósito fundamental es servir como estimadores de los resultados esperables en función del tipo de población que se atiende. Si, como es habitual, el *score* es una estimación individual del riesgo de muerte para cada caso, el riesgo de muerte de una serie de enfermos es la media de los riesgos individuales. La comparación entre la mortalidad observada y el riesgo promedio estimado es lo que dará idea del resultado en condiciones de ser comparado con otra serie. La manera más intuitiva de manejar estos datos es realizar el cociente entre la mortalidad observada y el riesgo medio de mortalidad estimada. Valores inferiores a 1 indican resultados mejores de los esperados y superiores a 1, peores a los esperados para ese sistema de predicción. Todo ello con las medidas estadísticas de dispersión correspondientes que permitan realizar comparaciones adecuadas.

En este uso es donde se plantean más interrogantes y dificultades a la hora de la interpretación de los *scores*. La pregunta habitual se refiere a qué *score* usar si queremos investigar la calidad de los resultados. La primera condición para su uso debe ser que sean «robustos» desde el punto de vista de su construcción. Otra condición recomendable es que el modelo se haya construido sobre bases de datos de pacientes que sean próximas o incluso mejor, a las que hayamos contribuido. En la actualidad considero que, en nuestro entorno, el *score* más recomendable es el EuroSCORE[10].

En resumen, este *score* se construyó sobre una base de datos europea del orden de 20.000 pacientes intervenidos durante el último trimestre de 1995. España contribuyó con más de 2.000 casos aportados por más de 20 servicios. Su funcionamiento ha sido suficientemente validado en nuestro entorno europeo con buena discriminación y calibración. Asimismo ha ocurrido al ser aplicado en poblaciones con diferencias demográficas importantes respecto a la europea, como son las norteamericanas. Por otro lado, aunque sea razonable la hipótesis de que este modelo no funcione bien al aplicarlo en pacientes operados sin ayuda de circulación extracorpórea, precisamente se ha documentado lo contrario[11]; es decir, una buena calibración y capacidad de discriminación al aplicarlo a pacientes coronarios operados sin circulación extracorpórea.

El modelo predictivo ideal sería el derivado de una base de datos extensa, que se actualice en el tiempo y que refleje los cambios que día a día se van introduciendo en la práctica clínica. Éste es el modelo de la STS que en la actualidad tiene una antigüedad aproximada de 10 años. En Europa se están sentando las bases para poder disponer de un modelo similar con la creación de una base de datos europea por parte de la European Association for Cardiothoracic Surgery que está en fase inicial.

La aplicación de los *scores* con el propósito de monitorizar la calidad presenta limitaciones serias y, por otra parte, está sujeta a errores de interpretación que son habituales.

En primer lugar, aunque estas herramientas son estadísticamente bastante sofisticadas, todavía no están en un grado de desarrollo que las haga infalibles. Estos *scores* están ponderando sólo una parte ínfima de las múltiples variables conocidas o desconocidas, del paciente o de la estructura asistencial, que influyen en el resultado final del proceso. Por esta razón, las conclusiones que se deriven de su aplicación han de ser cuidadosamente matizadas.

Otra limitación importante ya ha sido esbozada. Se trata de que los *scores* reflejan habitualmente una especie de «foto fija» de la práctica clínica en un momento determinado. El trabajo de García Fuster et al[9] es un ejemplo. El uso del EuroSCORE indica una progresión favorable de los resultados que se ha de atribuir, de acuerdo a la hipótesis de los autores, a una mejoría de la calidad prestada. Sin embargo, es muy posible que si se hubiera usado un modelo que reflejara la práctica actual, la visión de la progresión de los resultados no habría sido la misma.

Un error frecuente en el que se incurre se debe a lo que se puede llamar circularidad. Es frecuente que se utilicen *scores* como el EuroSCORE u otros, y se compare su capacidad de discriminación y calibración con un modelo derivado de la experiencia propia. Habitualmente, la capacidad de discriminación de los modelos propios es muy superior a la de los modelos externos. Esto, desde el punto de vista estadístico, es esperable. Pero desde el punto de vista del control de

la calidad, es inadecuado. Es fácil entender que un grupo con una mortalidad muy elevada podría desarrollar un modelo predictivo que tenga muy buena calibración y discriminación. Sin embargo, su uso no sería tolerable, pues sólo serviría para predecir unos resultados que podrían ser intrínsecamente inaceptables.

Por último, un problema frecuente es la interpretación errónea de lo que significa validación de un *score* determinado. Y no me refiero a los distintos tipos de validación metodológica a los que habría que someter a un modelo determinado. Me refiero a que cualquier uso de un *score* sobre una población habitualmente pequeña se suele denominar «validación» por parte de los autores. Es frecuente la conclusión de que determinado *score* no ha sido validado en nuestra experiencia.

En mi opinión, la validación de un determinado *score* significa la investigación de su calibración y de su capacidad de discriminación sobre una población determinada, pero con ciertas condiciones. La primera supone la aplicación estricta del *score,* sin aumentos artificiales del peso de las variables a todos y cada uno de los enfermos computables, sin excepción. La segunda condición, posiblemente la más difícil, consiste en evitar pérdidas en el cómputo del evento muerte según la definición que se utilice. La práctica diaria nos revela que, en el contexto de complicaciones, es fácil perder la pista de la información de un enfermo cronificado que finalmente fallece. Si el sistema de recogida de información no es perfecto, es fácil no computar adecuadamente a estos pacientes.

La última condición es especialmente importante en nuestro medio y se refiere al tamaño de la población para la validación del modelo. Como principio general se requiere[12] que la muestra para validación tiene que observar al menos 100 muertes. Esto quiere decir que si la mortalidad fuera del 5%, la población debería ser, por lo menos, de unos 2.000 pacientes. Muchos de los trabajos publicados con el objetivo de validar un determinado *score* manejan poblaciones llamativamente inferiores.

Si el intento de validación cumple estas condiciones y el modelo investigado no calibra y discrimina bien en nuestra serie, entraremos en el campo de las interpretaciones, que son especialmente complicadas. Se suelen dar dos escenarios típicos: sobrestimación o subestimación del riesgo. Si sobrestiman el riesgo frente a nuestros resultados, se suele concluir que nuestra práctica es correcta. Sin embargo, la inversa no se maneja igual. Es decir, si subestiman el riesgo frente a nuestros resultados reales se suele invocar a que están mal construidos o que están construidos sobre poblaciones muy distintas de las nuestras. Como suele suceder, ninguna de las posiciones es del todo correcta o incorrecta y la verdad está en una situación intermedia.

En resumen, los *scores* de riesgo son herramientas de extraordinaria utilidad, recomendables en la práctica habitual, no sólo quirúrgica, sino también en cualquier procedimiento con intención terapéutica, como la implantación de *stents.* Es necesario dimensionar cuál es el uso real que les queremos dar y, muy importante, tenemos que ser conscientes de las condiciones de uso que requieren y de las limitaciones y errores de interpretación que se les asocia.

BIBLIOGRAFÍA

1. Eagle KA, Guyton RA, Davidoff R, Edwards FH, Ewy GA, Gardner TJ, et al. ACC/AHA 2004 guideline update for coronary artery bypass graft surgery: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1999 Guidelines for Coronary Artery Bypass Graft Surgery). Circulation. 2004;110:e340-e437.
2. Kennedy JW, Kaiser GC, Fisher LD, Maynard C, Fritz JK, Myers W, et al. Multivariate discriminant analysis of the clinical and angiographic predictors of operative mortality from the Collaborative Study in Coronary Artery Surgery (CASS). J Thorac Cardiovasc Surg. 1980;80:876-87.
3. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. Ann Thorac Surg. 1988;45:348-9.
4. Cortina JM, Pérez de la Sota E, Rodríguez E, Molina L, Rufilanchas JJ. Escalas de valoración de riesgo en cirugía coronaria y su utilidad. Rev Esp Cardiol. 1998;51 Suppl 3:8-16.
5. Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. Ann Thorac Surg. 2004;77: 2232-7.
6. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation. 1989;79(6 Pt 2):I3-12.
7. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. Ann Thorac Surg. 2005;79:16-20.
8. Bernstein AD, Parsonnet V. Bedside estimation of risk as an aid for decision-making in cardiac surgery. Ann Thorac Surg. 2000; 69:823-8.
9. García Fuster R, Montero JA, Gil O, Hornero F, Cánovas S, Bueno M, et al. Tendencias en cirugía coronaria: cambios en el perfil del paciente quirúrgico. Rev Esp Cardiol. 2005;58:512-22.
10. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999;16:9-13.
11. Wu Y, Grunkemeier GL, Handy JR Jr. Coronary artery bypass grafting: are risk models developed from on-pump surgery valid for off-pump surgery? J Thorac Cardiovasc Surg. 2004;127:174-8.
12. Harrel FE. Regression models strategies. Berlin: Springer-Verlag; 2001.